PRINCIPLES AND METHODS OF STATISTICS

BY

ROBERT EMMET CHADDOCK
Professor of Statistics, Columbia University

HOUGHTON MIFFLIN COMPANY
BOSTON NEW YORK CHICAGO SAN FRANCISCO
The Riverside Press Cambridge

The Riverside Press
CAMBRIDGE · MASSACHUSETTS
PRINTED IN THE U.S.A.

TO

R. A. C.

PREFACE

This book is an introductory text for general use in colleges and universities. The object is not to contribute new results of statistical analysis, nor to advance the knowledge of the author's associates in this field. The purpose is rather to introduce the subject to beginners by the use of simple illustrative materials, to present the elementary principles of statistics in such manner as to appeal to the logical faculties of the reader, and to foster a healthy skepticism toward the results of quantitative investigations.

Elementary economics and social science are being taught to sophomores in college. Professional training begins in the last two years of the undergraduate course. Special courses include a varied background of facts which are intended to describe and to measure the phenomena of organized society. Many sources of information are in quantitative form, but figures do not necessarily tell the truth. Training in the scientific method of approach to a problem; acquaintance with the tools needed in his work; and familiarity with the methods of testing how reliable so-called facts really are — these are matters of fundamental importance to the student of the social sciences, theoretical and applied.

Technique and method in the physical sciences have been well developed and are applied in the experimental laboratory by the beginning student. In the social sciences many students are still denied the opportunity to acquire a knowledge of the methods of assembling data and of sifting evidence. They have not formed the habit of independent and cautious generalization from known facts. Statistics may serve a purpose for the social sciences similar to that which the experimental laboratory serves in the physical sciences. The possibilities for experimentation in society are very limited. The investigator must observe, record and compare phenomena, for the most part, as they are. Statistical method deals with mass data in their numerical aspects. It involves precise measurement, a careful record, intelligent and logical analysis and grouping, and discriminating judgment as to the relative significance of groups of facts.

Training in statistics should be regarded as discipline in scientific method. The undergraduate should be equipped with statistical method as a tool. Especially in the social sciences, where factors are many and varied and where human sympathies are keen, it is difficult to free judgments from the bias of desires and feelings. The scientific attitude of mind prompts the student to seek evidence which will appeal as true to minds other than his own. Quantitative data are objective and can be freed from bias more easily than subjective judgments.

The point of view from which the essential principles of quantitative analysis, description and comparison are approached in this treatise is logical and empirical rather than mathematical. A large number of technical and mathematical books on statistical methods have appeared in recent years. The author recognizes the value of these for the highly specialized student but doubts their utility for the beginner in the subject. The need at present is a clear understanding of fundamental principles in the treatment of numerical data.

Short methods have only a limited use in an elementary text since they may obscure for the beginner the meaning of the process which should be the all important consideration. Formulæ are regarded as short-hand statements of processes described and illustrated. The same problems are often carried forward from chapter to chapter, as the principles are developed, since by this plan the reader is freed from the necessity of becoming familiar with a new series of facts at a time when his attention should be concentrated upon the treatment of the materials.

Understanding in this subject results, for the most part, from doing. To find any value in the methods presented the student must make them a part of his experience and must be able to apply them with discrimination. Practice work, therefore, is the essential thing in acquiring statistical methods and in developing intelligence in their application. This book presents only a limited number of problems with detailed explanations of the methods applied to the data. In the author's opinion, it is the function of the teacher to select exercises for practice. In order to hold the interest of the student and to secure the best results the teacher will prefer to work out particular applications which interest himself and his students in specific fields. He will be able to select better materials for illustration than the author of a general text can possibly assemble for him. Besides, following a set of exercises prescribed by others emphasizes too strongly the mere routine of acquiring methods.

Most texts treat the methods described in Part III before those of Part II. This is the logical order in the collection and analysis of statistical data, and the teacher is at liberty to adopt this plan of presentation. The order of topics in this text has been adopted after trying out both plans in teaching, for the reason that very few have had experience in the actual collection of the original raw materials while most students have used to some extent existing sources of statistical data and possess a more or less extensive background of quantitative information. By treating at once methods of classifying and analyzing data already assembled the teacher builds upon familiar materials and is able to introduce the methods of Part III as the opportunity arises.

The scope of the book has been carefully limited to include only what can be covered in a year's elementary instruction and practice. The continuation of the subjects treated and the introduction of specialized methods are left to more advanced treatises.

The author has received invaluable suggestions from numerous text books and writers, for which he wishes to make acknowledgment and to express appreciation. An effort has been made to give full credit in footnotes for materials and ideas utilized in this text, but no such statements can adequately describe the debt we owe to the workers who have made their contributions to the methods of quantitative analysis now in general use.

Special acknowledgment is gladly given to Warren M. Persons and to the Harvard Committee on Economic Research for materials used in Chapter XIII; to Seymour L. Andrew, Chief Statistician of the American Telegraph and Telephone Company, for permission to use for illustration a section of their curve representing business conditions; to Willard C. Brinton and to the Society of Mechanical Engineers for permission to reprint the publication on Standards for Graphic Presentation; and to the authors of Bulletin 165, Agricultural Experiment Station, University of Illinois, for illustrative materials in Chapter XI.

I am especially grateful to my colleagues for their helpful criticisms and suggestions and for their generous spirit of coöperation: to Franklin H. Giddings, leader in the scientific study of society, for his sympathetic interest in the entire work and for a critical reading of Chapter III; to Henry L. Moore for reading the entire manuscript of Parts I and II and for discussing with me many of the topics presented; to Frank A. Ross, who has been closely associated with me in the work of statistical instruction for several years; to Wesley C. Mitchell for the use of materials and ideas from Bulletin 284 of the United States Bureau of Labor Statistics on Index Numbers, and for reading the manuscript of Chapter X; and to Russell G. Smith for assistance in revising Part III.

The author is indebted also to Mrs. Dorothy Reddy Van Der Veer for her excellent work in the final revision and preparation of the manuscript, and to Mrs. Charles A. Gulick for her assistance; to Mr. and Mrs. Roy E. Stryker for checking arithmetical computations and especially for their painstaking care in the preparation of the original drawings for the diagrams. He is under special obligation to Miss Estella T. Weeks for a discriminating reading of the proofs of the book.

Robert E. Chaddock

CONTENTS

PART I

INTRODUCTION — PRELIMINARY CONSIDERATIONS

PART II

CLASSIFICATION AND DESCRIPTION OF MASS DATA

PART III

THE GATHERING AND PRESENTATION OF STATISTICAL DATA

EDITOR'S INTRODUCTION

In spite of a widespread human interest in numerical facts, statistics, not long ago, was a subject which had only a small, though loyal, following. In recent years, the rapid progress that has been made in the prosecuting of statistical inquiries and the perfecting of statistical technique has been matched by an equally rapid general growth of interest in statistical studies. To this the increasing number of excellent introductory textbooks available to the student of statistics bears witness. It is no disparagement of other books to say that many teachers and students will find that Professor Chaddock's work will best meet their own particular needs.

It is the product not only of sound scholarship but also of long and conspicuously successful experience in the teaching of statistics. Throughout the book Professor Chaddock is the teacher as well as the statistician. This shows itself in the general plan and arrangement of the book, in the apportioning of space and of emphasis, and in the painstaking care given to every detail of exposition. It will be clear, furthermore, that Professor Chaddock has carefully appraised the needs, the purposes, and the initial equipment of the average student. For one thing, no mathematical knowledge is assumed, beyond the most elementary processes of algebra.

The book has another conspicuous merit. It treats of statistical methods as general rather than special tools. This, in my opinion, is as it should be. The student of vital statistics, for example, will be better equipped for his work if he knows something about the general uses, possibilities, and limitations of statistical method. So also, to take another example, with the student whose interest is primarily in the light which statistical analysis may throw upon business problems and policies. In particular, it is a mistake to put a fence around a narrow field, and dub it "economic statistics," or "social statistics," as the case may be. Students of economics are likely to profit quite as much by studying population statistics as by studying index numbers. And who will venture to say just what portions of the field that is coming to be called "business statistics" are not of importance to the student of economics? Professor Chaddock nowhere permits his interest in any one field of inquiry to obscure the importance of other uses of statistical method. In this respect, as in others, his book displays judgment and balance.

Whether statistics is a field of knowledge or a method is an old problem respecting which there was once much fruitless dispute, especially in German treatises. The dispute was fruitless because the issue was factitious. Statistics is not a field of knowledge so much as it is a particular type of knowledge. To organize this type of knowledge, to handle it effectively, calls for a special technique. The statistical method, then, is merely the utilizing of this special sort of knowledge, with the aid of this special technique, in any field of inquiry within which the method may be fruitful. The method, therefore, has a unity of its own, and this is reflected in Professor Chaddock's treatment of it.

The successful use of the method, however, requires a command of more than formulas and processes. Like any other scientific pursuit it calls on the one hand for the power of constructive imagination, and on the other hand for constant awareness of the necessity of guarding against mistaken inferences. In statistics the most dangerous sources of error are of a very elementary sort. There is just enough difference between statistical processes and the common types of reasoning that serve us well enough in the daily routine of life to make the statistical method a dangerous tool in the hands of the unwary. Professor Chaddock has done all that any writer could be expected to do, and more than most writers have done, to put students on their guard against the commonest forms of statistical error and to imbue them with the habit of critically scrutinizing their sources, their processes, and their inferences.

Throughout the book, moreover, he puts a needed emphasis upon analysis, as contrasted with the routine of statistical technique. I mean that he insists that the student should *understand* every element in the formulas he uses and every step in the processes he applies, and that he does all that he can to help the student toward such an understanding. It is here that Professor Chaddock has been most lavish in his expenditure of care and pains. The impatient student, very likely, will find some of this detail irksome. But the student with special faculty in mathematics or with a special aptitude for statistics will find nothing that will delay or hinder his progress. And for the great majority of students mastery of detail is the one sure road to a mastery of the subject.

Allyn A. Young

PRINCIPLES AND METHODS OF STATISTICS

PART I

INTRODUCTION

PRELIMINARY CONSIDERATIONS

The scientific study of any subject is a substitution of business-like ways of "making sure" about it for the lazy habit of "taking it for granted" and the worse habit of making irresponsible assertions about it. To make sure, it is necessary to have done with a careless "looking into it" and to undertake precise observations, many times repeated. It is necessary to make measurements and accountings, to substitute realistic thinking (an honest dealing with facts as they are) for wishful or fanciful thinking (a self-deceiving day-dreaming) and to carry on a systematic "checking up" . . . science is nothing more nor less than getting at facts, and trying to understand them.

Franklin H. Giddings

"Societal Variables," *The Journal of Social Forces*, March, 1923

PRINCIPLES AND METHODS OF STATISTICS

CHAPTER I

THE APPEAL TO FACTS

The basis of social action. An increasing number of communities are seeking exact information concerning themselves. Many classes of social and economic phenomena may be briefly and accurately described by quantitative statements which furnish precise measures of existing social conditions and significant changes from period to period. But more than description is sought. Careful comparisons of the facts of experience usually reveal relations of cause and effect, which make additions to our knowledge by explaining phenomena, not merely describing them.

The manner in which social and economic conditions affect individual and community welfare is measured in terms of family expenditures in relation to income, by industrial accident and mortality rates, by the decline of death-rates through the control of preventable disease, by the records of crime and dependency, by the length of the efficient working life, and by the changing standards of living of the population. These facts are as necessary for our enlightenment as those concerning the amount of exports and imports, the production of steel, the changes in commodity prices, and the volume of bank clearings.

The gathering and analysis of facts suggests the answers to many interesting and important questions. Is the Nation's production of food and raw materials keeping pace with the growth of population? Is the laborer better or worse off to-day than a decade ago, as measured by what he can afford to buy? Which occupations are dangerous to life and health, and why? What effect has social legislation on individual and social welfare? Are the wastes of community life growing less and the gains of coöperative activities growing greater? Is the immense increase in wealth really contributing to welfare as it should? Whether in estimating resources, or in measuring wastes, or in laying the foundation for intelligent public opinion concerning what is necessary for social and economic progress, quantitative measurements play a most essential part.

Social legislation based on facts. More and more frequently exact knowledge concerning conditions precedes the enactment of laws designed to modify them. The Government has been extending its activities constantly into fields where investigation is a necessary preliminary to wise action. Collection of facts concerning hazards in modern employments and a growing realization of the economic and social consequences of neglect, have been followed by agitation for better protection of workmen, by preventive methods or by social insurance. Before either policy could be adopted careful investigation was needed to determine the causes of industrial accidents and to place the responsibility upon specific processes or trades; and to determine what burden insurance must meet. During the last decade almost all our commonwealths have enacted compensation laws.

Observation has shown the comparatively low pay of women and minors in certain industries and the resulting low standard of living with its social consequences. It has been proposed to protect this class of wage-earners in a special manner by minimum-wage laws. But before the legislature and the administrative authority can act intelligently, it is necessary not only to know what wages are being paid in specific industries but also what wage is required in order to meet the reasonable needs of the worker. These are questions of fact and must be determined by inquiry.

Vital statistics in the service of sanitary science and health administration. Modern sanitary science owes its existence to the registration of deaths and their causes. Records of death and sickness constitute the bookkeeping of the public health movement. They direct the activities of the sanitary expert as chart and compass guide the navigator. Vital statistics furnish a definite measure of the value of sanitary improvement and of the progress of preventive medicine, as the following facts indicate:

TYPHOID FEVER IN PITTSBURGH, 1907–1911 [a]

YEAR	NUMBER OF CASES	NUMBER OF DEATHS	DEATH-RATE PER 100,000 POPULATION
1907	4,514	502	130.8
1908	1,833	255	46.6
1909	955	130	24.6
1910	998	149	27.7
1911	768	140	25.9

[a] From the Annual Report of the Department of Health, Pittsburgh, Pennsylvania, February 1, 1911, to January 31, 1912, p. 56.

The annual average death-rate 1900 to 1907 was almost exactly the same as that for the single year 1907. In 1920 the rate was 5.6 deaths per 100,000 of the population.

During 1908 filtered water was first supplied by the city at great expense. The death-rate from typhoid was only one third that of the preceding year, representing the saving of several hundred lives. For several years a part of the city continued to use unfiltered water. In 1911 one quarter of the population which used the unfiltered water supply contributed over half of the total deaths from typhoid. The death-rate of this section was four times as high as was that using filtered water.

Vital statistics aid in the efficient and economical administration of a health department. Large cities spend millions in the protection of health. The division of records and research should relate health facts to different sanitary conditions and the housing of the population, to nationality, occupation and the standard of living, in order to enable the health executive to concentrate his efforts at the weak spots in the city's health defense. Only in this manner will the expenditures yield the largest returns in saving lives and promoting health. Finally, the health official must use facts to arouse public opinion and to gain financial support for the future development of health policies.

Statistics and court decisions. Many laws are enacted to-day with the intention of protecting the health and welfare of the working population, especially women and minors. The courts decide whether a law is a valid exercise of the police power of the state, or is in conflict with constitutional guarantees. *Their decisions turn upon questions of fact.* The issue may be the proper length of the working day for women in certain specified industries, or whether or not women should be permitted to work at certain processes dangerous to health. If it can be shown that a working day longer than the law authorizes or that the particular processes are injurious to the worker's health or to the exercise of her functions as the mother of the next generation, the action of the legislature will probably be upheld by the court. Vital statistics are needed to measure the effect of the industrial process upon the health of the worker. In the past there has been a great dearth of occupational health statistics in the United States.

Statistics and modern business. Since the business unit has increased in size, since business organization has grown more complex, and since markets have developed over wider areas, the factors which make for success or failure have increased in number and have become more difficult to estimate without systems of precise record. A guess based on the experience of individuals, even if it be expert, is not sufficient.

The functions of the bookkeeper have been greatly enlarged. Formerly he kept track of a few general items of business, moneys received and paid, accounts receivable, the payrolls, the bank balance, and the like. To-day, however, mechanical devices have relieved him of much of the drudgery and he records a dozen items pertaining to the business for every one he recorded before. His mind is free to plan new records and new methods of analysis which will increase the effectiveness of business organization. In modern business, bookkeeping has developed into a science of accounting.

Furthermore, business statistics include much more than the financial aspects of the concern. Almost every one connected with the business keeps some sort of record. Streams of data from branch factories and salesmen on the road flow into the central office. These are consolidated and interpreted and the executives know the current operations compared with the facts of last month and last year. Departments learn their relation to other departments, and the complex business moves in harmony, with a minimum of waste. The best methods of work, the correct standards for materials, the wastes of labor turnover, the ability of men and their fitness for specific work, are all determined on the basis of recorded experience.

More recently employers have begun to record certain human factors in business. Compensation laws compel remuneration for injury to workmen for industrial accidents. It is a matter of economy, therefore, to prevent accidents. Careful records are kept, dangerous processes are located, and ignorant workmen are instructed. Absence on account of illness or other causes is a disorganizing factor in modern business, and, since it interferes with operation and causes loss, it should be prevented as far as possible. Records reveal that in different industries and in different processes within the same factory illness among the workmen varies widely. By control of industrial conditions much disability can be prevented. Many business concerns are doing this and thereby increasing their efficiency. A knowledge of these facts has led some employers to provide a physician to examine their employees at the time of hiring and at intervals afterward, in the hope of keeping them well.

Facts other than those pertaining to the internal affairs of the particular business are useful to the management. Cycles of prosperity and depression in general business, movements in commodity prices, general credit conditions, marketing possibilities, labor and trade movements are measured by series of quantitative data. These and many other factors constitute the environment of any particular business enterprise, and condition its success. They should be known and related to the internal facts and policies of the individual concern.

The modification of old theories by new facts. The accumulating records of experience are valuable checks on the accuracy of assumed premises and the logical deductions drawn from these premises. Theories and hypotheses which do not square with the facts of life are being criticized and modified. Only recently have the social sciences reached the stage where the necessary facts are being assembled, classified, analyzed, and compared, with the definite intention of checking old hypotheses and laying a firmer foundation for scientific conclusions.

The older theory that vice and crime were chiefly the results of individual choices and personal responsibility, grew out of the assumption *that the person was entirely free to choose.* This theory, when applied to modern associated activities in city and factory proves inadequate as an explanation in the face of more knowledge gleaned from the recorded experiences of men. Vice and crime have their individual aspects, but to-day they are being carefully analyzed also as social products. The lack of proper play and recreation facilities, the results of the failure to separate juvenile delinquents from hardened criminals, the employment of children and young persons in situations dangerous to morals, the neglect of the mentally defective who themselves lack self-control, the needlessly early death of parents on account of preventable accidents and sickness, the results of defective education for boys and girls, and low standards of living, are all being considered as possible causes. The relation of these conditions to individual character and conduct is growing clearer as the facts accumulate. The frank acceptance of community responsibility for the continuance of conditions beyond the control of the individual characterizes modern social movements.

No doubt there are individual causes of poverty, but one who attempts to account for the full extent of poverty on the theory of individual responsibility offers only a partial explanation and no basis for elimination. Underlying causes are left as before. Recent investigations seek to analyze the various factors involved. Thousands are killed and scores of thousands are injured in industry every year. A single disease, tuberculosis, causes one hundred thousand deaths annually in the United States, chiefly at the most productive period of life when family responsibility is heaviest and earning power is greatest. For every death from tuberculosis there are probably at least five persons who are ill from the disease and whose efficiency and earning capacity are more or less impaired. These and many other preventable causes of illness, disability and death undermine standards of living. Mothers become chief breadwinners, children leave school early and are deprived of the opportunity for training which would make them better income providers as adult workers. Economy requires more restricted living conditions. Low earning power, ill health and consequent low wages unite to condemn the family to perpetual poverty. It is agreed that much of the loss of life and health is preventable, but the individual cannot do it alone. Social action is required, both to prevent needless wastes and, where prevention is not possible, to diffuse the burdens by schemes of social insurance.

Much employment in our present industrial organization is seasonal in character. The workers experience long periods of slack work. They suffer the consequences of intermittent earnings insufficient for their needs. Their standard of living is endangered through no fault of the individual worker. A closer analysis of the classes of the unemployed reveals some who are out of work because they have lost their vigor and efficiency through exposure to bad living or working conditions, others because they have become unreliable through frequent shifting from job to job, or on account of bad habits, or because they are mentally defective. These classes of the unemployable experience the most hopeless poverty, and for much of it they are not responsible as individuals.

From the accumulation of such facts as have been cited and many others of similar character, the causes of poverty and misery are shown to be social as well as personal. It becomes the responsibility of the community to remove these handicaps and to give to the individual a chance. Then the individual may be held accountable for the use he makes of his opportunities.

Theories of government. Many have believed that the government is best which governs least. In the light of the results of the *laissez-faire* policy this theory has been gradually modified. The freedom of the individual is being limited in the interest of the many. The principle of enlightened self-interest, held by Adam Smith and his followers to be a *sufficient* guide for the activities of men in organized society, was proved to be socially inadequate by the evidence recorded in the reports of investigating commissions disclosing the exploitation of women and minors in English factories during the early part of the nineteenth century. Employers were following the motive of selfish gain. The story of industry in the United States is no different. It becomes the function of government to establish standards of protection for the workers in industry, and to compel conformity to these standards in order that the pursuit of profits in business competition may not be permitted to destroy or impair the lives and health of the masses who toil.

In public service utilities, as the street railways, monopoly is frequently the inevitable and economic form of business organization. To protect the consumer against the undue use of the power conferred by monopoly, government interferes more or less with the conduct of private business. Through the Interstate Commerce Commission an attempt is made to regulate conditions of railway transportation. Public utilities commissions in our cities supervise the quality of service. It is decided that more than a specified price cannot be charged for gas, so long as this price enables the producing company to pay a reasonable interest on its capital or investment. The owner of a city lot, in the interest of health, is forbidden to cover the entire area with a building designed for dwelling purposes. The adoption of policies of control and regulation by government agencies, in these and many similar fields, has followed the more or less careful record and analysis of the results of experience under the opposite policy of non-interference.

SUMMARY

The purpose of the preceding discussion has been to emphasize the growing importance of exact information as the basis for generalizations, for rational explanations of phenomena and for policies of action. The tendency to-day is to base both thought and activity upon broad foundations by the investigation of present conditions and past experience. Statistical records play a more and more important part in reducing our knowledge to a precise and objective form in estimated, counted or measured units. These data may be recorded from period to period, from place to place, or in varying magnitude. They may be classified, analyzed and compared. Behavior of persons and groups may be characterized in an objective and impersonal manner by appropriate statistical methods.

[Introductory Readings and References are given at the close of Part I, Chapter III.]

CHAPTER II

MISUSES OF STATISTICAL DATA

The importance of method. Are not statistics merely the results of enumeration and measurements, or of making estimates, where exact counting or measurement is not possible, set forth in the form of tables and graphic devices? Even if these operations were our sole interest in statistics, they are not simple to perform with accuracy and completeness. Even if the unit to be counted or measured is clearly understood, counting is not easy. Several readers, asked to count rapidly the capital letters on five pages of text would not agree in their results, yet there must be a fixed number. What is a statistical fact? The geodetic survey does not trust one very exact measurement of the length of the base line, for purposes of triangulation. A single observation proves less reliable than the average of several measurements. The wage-earner does not remember how many days he lost on account of illness during the past year. The housewife does not keep an account book and, therefore, does not know how much she spent for food during the past month. Evidently statistical facts are not merely figures.

When the term *statistics* refers to the numerical data which constitute the raw materials of an investigation the plural verb should be used, as in the preceding paragraph. But statistics *is* much more than measurements, countings, or estimates. In fact the original data are trustworthy only when approved methods of collection are employed. *To emphasize the unity of the principles with which statistics is concerned as a scientific method, the singular verb is used.* The investigator is guided by his knowledge of statistical methods in defining clearly the unit to be counted; in devising ways of eliminating errors of observation and personal bias, and of estimating how accurate measurements really are. Statistics has to do with methods of presenting in abbreviated and classified form the facts concerning large groups of phenomena which are too complex for simple observation. It is impossible, from a glance at the detailed payrolls of two factories employing thousands of men and women, to contrast with accuracy the wage conditions, or to say with assurance that the laborers as a group are better paid in one factory than in the other. Finally, methods exist for discovering and measuring relationships between various series of data. It may be found that the death-rate of children under one year of age varies as the income of the family changes.

Observation of a few cases often leads to hasty and mistaken conclusions. Are there ways of guarding against these unwarranted generalizations? The use of statistical methods is limited to the numerical aspects of a problem. Good judgment and common sense are always needed to bring these into proper relationship with the non-numerical.

Sound facts are frequently made the basis for unwarranted inferences. This chapter is designed neither to present technical definitions nor to discuss statistical difficulties, but to show in concrete terms that statistics is much more than mere numbers; that the methods of using facts are of supreme importance to student and research worker; and that the critical attitude toward conclusions from numerical data is the only safeguard. Data may look convincing because they are in precise form and seem to be a final statement of truth. But figures do not bear on their face the stamp of credibility. They may not mean what some one makes them appear to mean. Illustration will make this point of view clear.

COMPARISON OF NON-COMPARABLE DATA

Statistical data are especially useful for making significant comparisons. *The most necessary warning is to be sure that things compared by means of quantitative measures are really comparable.* Neglect of this principle has been the cause of many faulty inferences.

Criminality among the foreign-born. The census authorities, from the statistics of prisoners in institutions in 1890, drew the conclusion that *the tendency toward criminality among the foreign-born was twice as great as among the native-born population.* It was found that there were 1768 prisoners in institutions per million of the total foreign-born population of the United States as compared with only 898 per million of the total native population. But a group of a million foreign-born cannot be compared with a like number of native-born as to the occurrence of crime. Prisoners are recruited mainly from adults, and the proportion of adults among a given number of foreign-born is much greater than among the same number of native-born. The native-born group includes a larger proportion of children and old persons who do not contribute to the number of prisoners. In other words, the million foreign-born and the million native-born were not comparable as to potential criminals. A grave injustice was done the foreign group by this conclusion. If we compare for each of the following groups the number of *male prisoners* in institutions in 1890 *per million males of voting age* in the population, a very different result appears.[1]

Native white of native parentage	3395
Native white of foreign parentage	5886
Foreign-born white	3270

[1] J. R. Commons: *Races and Immigrants in America*, pp. 168–69.

Death-rate from disease among American soldiers in the Philippines. In reply to the adverse comment which had arisen, the report of the Secretary of War, for 1899, discussed the death-rate from disease among the soldiers in the Philippines. The report compared the annual death-rate among the soldiers, 17.2 per thousand, with the rates among the general populations of Washington, D.C., and of Boston. Since the rates appeared to be about the same, it was maintained that the soldier rate should not be considered excessive. But one thousand soldiers are not comparable with one thousand of the general city population, with its large proportion of very young and old among whom the death-rate is always high. Soldiers form a selected group, both because they are all in the middle age period where death-rates are low and because they are examined physically before entrance into the service. Comparisons of this kind are in no way scientific, and obviously are invalid.

Comparison of coal-mine accidents in different countries. Comparisons of the annual number of fatal accidents per thousand workers employed in the mines have been made for the leading countries. In this manner the different countries are ranked in the order of relative hazards in the coal industry and their progress in protecting workmen against the dangers of coal mining. But these comparisons are not fair because no account has been taken of the number of days the mines were in operation in the different countries. For instance, on the average the coal mines of the United States operate fewer days than do those of Europe and, therefore, the former expose their workers fewer times during a year to specific dangers, irrespective of protective devices which may be employed. This makes the accident rate for the United States appear more favorable than it should. The best practice now reduces each country's accidents to a common three hundred working-day basis, regardless of the actual number of days in operation. This device in statistical method, which has the effect of stating the average number of accidents per man per day, makes the data for the different countries comparable in respect to the number of days of exposure to accident during the year.

The measurement of unemployment. Certain States — for example, Massachusetts and New York — through their labor departments, collect facts as to the numbers and proportions of trade-union members unemployed each month. From these figures can we estimate the total numbers unemployed in general industry? The answer turns upon whether the amount of unemployment in union trades is typical of the general employment situation.

Incomparability of data in the statistical sources of different States. We wish to compare two States as to their protection of workmen against industrial accidents in a specific industry. We use the annual accident rate per thousand exposed workers in each State as the basis for our comparison. But one State requires all accidents causing any loss of time to be reported, while the other requires only those causing the worker to be absent for one week or more. Evidently total accidents as recorded in the two States do not reflect the actual situation for comparative purposes. The figures are not comparable, although they seem to measure the same phenomenon. The comparability of many types of statistical data collected by States and cities is destroyed by similar lack of uniformity. *It is never safe for purposes of comparison to accept published statistics at their face value without careful scrutiny of their limitations.*

Tuberculosis among men and women in the garment trades. Voluntary physical examinations to detect the presence of tuberculosis among garment workers in New York City are carried on at the Union Health Center. From the records it appears that a distinctly larger proportion of male than female workers contract the disease. But the two groups, male and female, in the garment trades are not comparable in respect to their ages, and this is an important factor in the incidence of tuberculosis. The average age of the females is kept low because they are constantly leaving the trade and are being replaced by young women entering it. Conclusions concerning the relative susceptibility of males and females in this trade are, therefore, likely to be misleading, unless care is taken to compare only similar age groups of both sexes.

Comparative size of the family in two generations. A bulletin, prepared by the Massachusetts Bureau of Statistics of Labor, based upon the returns of the State census of 1905, was entitled "Comparative Maternity." The schedule used for females asked for information as to the number of children born to each living mother in the State, and also asked each of these women to state the number of children born to her mother. From these records the average number of children born to mothers of two generations were computed and compared. Startling results were obtained. For instance, it was stated that "while the native-born mothers had an average of 2.77 children, their own mothers had an average of 6.47." This comparison made it seem as if there had been a great decline in the size of the families of the recent generation (1905) as compared with the preceding generation. But this comparison was obviously invalid because many of the mothers living at the census date had not completed the child-bearing period and their families would continue to grow.

WRONG USES OF PERCENTAGES

Need for both absolute numbers and percentages. A short time after Johns Hopkins University had opened certain courses in the University to women, it was reported that thirty-three and one third per cent of the women students had married into the faculty of the institution. Of course the important information was the number of women students. There were only three. *When dealing with a small number of cases, the use of percentages alone leads to wrong impressions.* In these cases either percentages should not be used at all or the numbers upon which they are based should accompany the percentages.

Percentages as proportions of a total one hundred. The following table is from page 61 of the *Weekly Bulletin* of the New York City Department of Health, February 20, 1915:

Percentages of Total Deaths under One Year, by Chief Causes

	1907	1914
Diarrhœal diseases	31	22
Respiratory diseases	21	22
Congenital debility	32	42
Contagious diseases	4	4
All other causes	12	10
Total	100	100

In describing the significance of the figures, the statement was made that "the respiratory diseases show a slight increase. The deaths from congenital debility show a marked increase from 32 per cent to 42 per cent of the total deaths." This manner of statement, together with the table of percentages, is likely to convey a wrong impression — that the absolute number of deaths from congenital causes has increased. As a matter of fact the death-rate under one year of age in 1907 was 144 per thousand births, while in 1914 it was only 95 per thousand births, a remarkable decline. This was the period of the establishment of the infant milk station and baby health center. In Greater New York there was organized in 1911 a united campaign for the decrease of infant mortality. The success of this movement was most marked in reducing deaths from diarrhœal diseases, *but the absolute number of deaths declined in each of the main causes stated in the table.* The respiratory and congenital deaths were not reduced as rapidly. Therefore, in 1914 the diarrhœal diseases constituted a much smaller proportion of the *smaller total infant deaths* than in 1907, and it follows that the other causes must constitute a larger proportion, because the total percentage was 100 in both years. Because percentages are proportions of a total 100, a decrease in one involves a corresponding increase in the others.

Averaging or combining percentages. A city of 100,000 population has 20 per cent foreign-born, a second city of 500,000 has 30 per cent, and a third city of 1,000,000 has 40 per cent. What is the proportion of foreign-born in the three cities combined? A simple average of the percentages

$$\frac{20+30+40}{3} = 30$$

is not correct, as the detailed computation indicates.

100,000 times 20 per cent =	20,000 foreign-born
500,000 times 30 per cent =	150,000 foreign-born
1,000,000 times 40 per cent =	400,000 foreign-born
1,600,000	570,000

which equals 35.6 per cent foreign-born. It is clear that 35.6 per cent of the population of the three cities combined are foreign-born. *In combining percentages of different sized aggregates into a single value, it is necessary to weight each percentage by the size of the aggregate to which it refers.* In this example the procedure amounts to taking the total foreign-born in the three cities and computing this total as a proportion of the total population of the cities.

One of the statistical problems which confronted the Lane Railroad Wage Commission in 1918 was a study of the adjustment of wages in transportation to meet changes in the level of prices. The specific task was to secure a single figure which would measure percentage change for all the items of the family budget collectively. On page 82 of the Report of this commission to the Director-General of Railroads in 1918 the per cent increase in the cost of various items of the family budget from January 1, 1916, to January 1, 1918, are given:

Food	52 per cent
Rent	10 per cent
Clothing	44 per cent
Fuel and light	31 per cent
Sundries	35 per cent

Food and clothing showed the largest increases in cost, but these two items are of very unequal importance in the total expenditure of an average family. Food requires somewhat less than one half of the total family expense, whereas clothing requires only about one seventh of the total.

It was necessary to combine these percentages into a single figure to indicate how much wages would have to be increased to meet the changes in the cost of living. A simple average

$$\frac{52+10+44+31+35}{5} = 34.4 \text{ per cent}$$

would give equal importance to each item of the budget. But these increases in prices affect the family expenditure for various items in proportion to the percentage of the total expenditures of the family devoted to each item. On page 87 of the Report these facts are given for families spending a specific amount — for instance, $600 to $1000 per year. To secure a single figure measuring change in the total cost of living we may combine the various items.

	Per cent Increase in Prices (1)	Per cent of Family Expenditure (2)	Product of (1) and (2) (3)
Food	52	42	2184
Rent	10	20	200
Clothing	44	14	616
Fuel and light	31	7	217
Sundries	35	17	595
		100	3812

$$\text{Average change in cost of living} = \frac{3812}{100} = 38 \text{ per cent.}$$

This method of combining the various percentage items, by using column (2) as weights, gives a more accurate picture of the actual change in the cost of the entire family budget than the simple average.

NEGLECT OF IMPORTANT FACTORS

Frequently an incomplete or erroneous explanation results from presenting quantitative evidence concerning only one factor in a complex of many factors. A part of the truth may be generalized into a complete explanation of cause and effect. Other considerations even more important may be neglected entirely. A well-known humorist has declared that statistics is the art of stating precisely what we do not know.

Smoking and failure in college. The annual report of a Western college presents the startling conclusion that smoking is the cause of failures. What evidence is used as the basis for this generalization? The male students were classified into three groups — non-smokers, moderate smokers, and excessive smokers. An investigation was made of the marks and proportion of failures among each of the three groups with these results.

	Non-Smokers	Moderate	Excessive
Number of students investigated	111	35	18
Average work for year	85.2 per cent	73.3 per cent	59.7 per cent
Proportion of failures	3.2 per cent	14.1 per cent	24.1 per cent

The association of smoking and a high percentage of failures is clear from this evidence, but the assertion that the one is the cause of the other may not be warranted. The men who smoked excessively in that college were probably those who valued other things more than marks. They may have indulged in social activities; they probably were prominent supporters of athletics; they were, in most cases, the fellows who did not come to college to study. Smoking was one of the ways of passing the time agreeably. In any case the number of excessive smokers was small. Other hypotheses than smoking might account for their low grades and failures. *Half-truths or quarter-truths should not be accepted as whole-truths simply because supported by evidence in numerical form.*

Physical defects and repeaters in school. An investigator attempts to find out why so many children in our public elementary schools are behind their proper grades and are repeating work. The records of the physical examinations of these children reveal many defects of eyes, ears, nose and throat, and teeth. The investigator may conclude that physical defects are the cause of the retardation of pupils and their inability to keep up with their classes. This generalization would not be justified unless first he had investigated other possible factors in the problem, for instance, the frequency of transfer of pupils from one school to another, the ability of the parents to speak English, and the enforcement of the truancy regulations.

When several factors are involved in producing a specific result, conclusions should not be drawn from the measurement of only one factor. The attitude of mind should not be that of the debater who counts on stating his case in the strongest possible terms, allowing his opponent to check up and refute by such facts and arguments as he can find. The effort of the scientific investigator should be to weigh and to measure every known factor in the problem before hazarding a conclusion.

THE BIASING INFLUENCE OF PREVIOUS CONVICTIONS

We live amid a wilderness of recorded data. Enthusiasts, exalted by visions of a new order, or the self-interested, impressed with the need for a defense of their views, seize eagerly upon the figures to be found in reports called statistics, and appropriate such supposed facts as suit their purposes. Then they proclaim their version of the facts as truth. Therefore the impression has been created, not without justification, that it is possible to prove almost anything by statistics. It has been declared that there are two ways of deceiving people, that is by perjury, and by statistics. The victim of deception is not infrequently the person who collects and uses the data. He who takes the point of view of an advo·

cate is on dangerous ground. It is for this very reason that a knowledge of methods of gathering, analyzing and comparing data are important.

Labor's share as described by the agitator. A writer in *The Survey* of February 8, 1913, pages 653–54, in discussing the fixing of wages by law, says, "The Report of the United States Bureau of Commerce and Labor (*sic*) for 1910 states how labor received only 20 per cent of the value of the product which it serves to create." The Report referred to makes no such statement nor could such an inference be drawn from the data which the Report contains. The Census of Manufactures, covering statistics for the year 1909, reported the total value of products in manufacturing industries as more than twenty billions of dollars. Wages and salaries in these same industries were estimated at about four billions, which is twenty per cent of the value of the finished products. These simple facts were made the basis for the inference that labor received only twenty per cent of the value of the products *which it served to create.*

But, even with a limited knowledge of industrial processes, it is obvious that laborers in manufacturing do not create the entire value of the products ready for the market. The laborer in the factory does not create the raw materials; it is the laborer on the farm and in the mine and forest who produces the supply of raw materials, to say nothing about the part which Nature plays. The report itself stated the value of these materials entering into the processes of manufacture as over twelve billions of dollars. Surely the laborers in manufacturing had no part in creating this value. Therefore, subtracting twelve billions from twenty billions *leaves about eight billions as the value added by fabricating processes.* Of this, labor received about four billions, or fifty per cent instead of the alleged twenty per cent, and this estimate makes no allowance for replacement of capital, equipment, interest charges, taxes, and other expenses which must be deducted before profits emerge.

The article suggests as a remedy for this exploitation of labor, that laborers be awarded by law a minimum of thirty-three and one third per cent of the value of the product, whereas they were already getting, according to all the facts in the Census Report quoted, more than fifty per cent.[1] The writer had used the facts in a manner not only to misinform other people, but also to deceive himself into making a perfectly ridiculous proposal for the supposed good of the working classes. He had welcomed the statistics to support a conviction, and had used them without discrimination or judgment.

In 1896, S. N. D. North, a former Director of the Federal Census,

[1] A recent authoritative discussion of this subject, with the supporting quantitative evidence, will be found in *Income in the United States*, vols. I and II, National Bureau of Economic Research, New York City.

pointed out similar erroneous deductions from the statistics of manufactures. At that time he criticized writers for drawing conclusions "as to the relative shares of labor and capital in the joint product, which are a travesty upon the facts." He further declared that such inferences formed the basis for socialistic teachings by those who wished to use the facts to prove that labor was being robbed of an equitable share of the product.

Vaccination and its enemies. In the *New York Evening Sun*, of May 4, 1914, appeared an open letter from the treasurer of the Anti-Vaccination League of America, setting forth evidence to show that to-day compulsory vaccination is unnecessary and ineffective for the prevention of epidemics. He cited the following statistics for England and Wales for the years 1905 through 1910 from the Annual Reports of the Registrar-General of England — an excellent source.

Total deaths from smallpox for six years, 1905–1910	199
Total deaths from vaccination for six years, 1905–1910	99
Deaths from smallpox, under five years of age, 1905–1910	26
Deaths from vaccination, under five years of age, 1905–1910	98

From these data he concludes that vaccination is not a protection to the public health and should not be compulsory because "this fearful and latest published record of registered figures shows that the total deaths from vaccination in England are about half as much as the total deaths from smallpox for all ages; but if we compare the deaths under five years for both causes we will find the astounding result staring us in the face that the actual deaths from vaccination among children are nearly four times the actual deaths from smallpox."

The figures are from a reliable source, but the conclusion fitted a previous conviction and was not in accord with a common-sense interpretation of the facts. Why were there so few deaths in six years from this disease, once feared as a scourge? It is scarcely necessary to suggest that vaccination was largely responsible. Of what significance are ninety-nine deaths in six years in a population of millions, compared with the lives which might have been sacrificed by the epidemics of smallpox if vaccination had been neglected? This is such a glaring misuse of statistical data that it seems incredible that the reader of the daily press should be compelled to guard himself against it.

AN ILLUSTRATION FROM AN OFFICIAL SOURCE

In 1891 the United States Senate directed its Finance Committee to investigate the effect of the tariff on wages and prices. Its Report has been widely quoted as the authoritative investigation in the United States on wage changes previous to 1890, the date which marks the beginning of the series of index numbers constructed by the Bureau of Labor Statistics. The data on wages were collected by the late Carroll D. Wright who had been in charge of the Census of 1890, but he was not responsible for the analysis of the facts or the conclusions drawn from them. *The final Report stated that wages had increased sixty per cent in the period from 1860 to 1891.* The validity of this conclusion can be determined only after an examination of the scope and method of the Report.

1. The data were secured from the paysheets of eighty-eight establishments in twenty-one different industries. This is a reliable method of securing wage facts. The wage-earners were classified into over five hundred subdivisions of industry. For example, the workers in the brewing industry were divided into five groups — master brewer, foremen, coopers, teamsters, and laborers.

2. After the facts had been assembled the first task was to measure the changes in wages over the period, for each subdivision of a given industry. We shall use the brewing business as an illustration of the method employed, because it has a small number of subdivisions. The procedure was to secure the wages of all the foremen, for example, in the business for January and July, 1860, and for the same months in all the successive years inclusive of 1891. The average wage for 1860 was made the base and called 100 per cent. The average wage for 1861 was represented as a percentage of the 1860 wage and called the index for 1861. In illustration, if the average for foremen was $5.00 per day in 1860 and had advanced to $6.00 in 1861, the index for 1861 would be 120. The index for each successive year was computed in the same manner, basing all percentages on the 1860 wage. In this fashion percentages were calculated for the other subdivisions of the brewing business, for each year of the thirty-year period, all based upon 1860 wages. These percentage indexes, year by year, showed the changes in wages.

3. The next step was to obtain an average percentage change in wages for the entire brewing business, year by year, from 1860 to 1891. The facts for the last year of the period are presented in the table.

PERCENTAGE CHANGES IN WAGES FOR THE BREWING INDUSTRY, 1860 TO 1891 [a]

SUBDIVISION	1860	1891
Master	100	375.0
Foremen	100	195.1
Coopers	100	185.4
Teamsters	100	145.8
Laborers	100	222.4
Brewing industry	100	$\frac{1123.7}{5} = 224.7$

[a] Senate Document 1394, 52d Congress, 2d Session, vol. I, pp. 112 and 173; and vol. II, pp. 312 *et seq.*

Wages in the industry in 1891 were described as 225 per cent of what they were in 1860. For every dollar which the workers received at the earlier period they were receiving $2.25 in 1891. A simple average of the percentage increases for the five subdivisions was calculated without reference to the varying numbers of workers employed in each subdivision. In other words, each subdivision was given equal weight in the final average, 225 per cent. By a similar procedure an average percentage was secured for each of the industries.

4. Finally, a single percentage was calculated for the totality of industry, by a simple average of the percentage changes for the individual industries. In this case also equal weight was accorded to each industry regardless of the varying numbers of workers employed and affected by the change in wage.

Criticism of the method and scope. 1. The most important criticism is directed against allowing equal weight in the average for the entire industry to each subdivision of the business. *The entire brewing industry was represented by one firm* located in New York, in which there was only one master brewer. His wages showed by far the largest increase, 275 per cent. The method of simple average used in the Report allowed equal weight to the master's large increase and to the smaller increases of the other groups. The result was that the master's large increase unduly raised the average for all and gave an erroneous impression as to the wage increase in the industry.

2. This criticism becomes more important when we learn that the master brewer was not the only one-man series in the investigation. Half of the industries were represented by only one establishment each,[1] a very narrow basis for inferences concerning wage conditions in the entire industry. It is difficult to find one firm that is representative of an entire industry.

3. The importance assigned to each subdivision of an industry remained the same throughout the thirty-year period. The fact is well known that the relative numbers employed in various branches of an industry change in successive years, due to the introduction of improved machinery and new processes. Therefore, the weights for the various subdivisions should be changed from year to year in computing the average percentage change for the entire industry. If the numbers employed in each subdivision had been used as weights, this shifting within the industry would have been recognized and allowed for from year to year.

4. The number of workers investigated in each subdivision of the in-

[1] *Op. cit.*, vol. I, pp. 111 *et seq.*, Table 37; vol. II, Table 12, for details of each establishment.

dustry formed a small proportion of the total engaged in that type of work in the entire country. Half of the industries were represented by only one establishment each, and the sample selected was not always typical.

5. The data represented the manufacturing and transportation industries located largely in the north and east of the country. No attempt was made to study farm wages, although at the period 1860 through 1891, the country was mainly agricultural. Therefore, this report on wages in the United States can lay no claim to completeness and its conclusions have very definite limitations in scope.

SUMMARY

In presenting these examples of wrong uses of statistics, no attempt has been made to question the accuracy with which the data were collected. The purpose has been to center the attention of the student upon the importance of methods of handling quantitative data. Statistics is more than figures, and the equipment required is more than adding machine and calculator. To discriminate between the true and the false inference drawn from figures requires the exercise of the critical faculties and a training in scientific methods.

The student of statistical methods and of published results is urged to cultivate a critical point of view toward quantitative data and their uses. It is suggested that he keep a sharp lookout for examples of wrong uses and procedures, not only in the reading required in various college courses, but in outside reading in newspapers, magazines, and scientific journals. Effort should be made to characterize and classify each instance under specific types of misuse, by attempting to explain in each case why an indefensible inference has been drawn or a wrong procedure has been employed.

[Introductory Readings and References are given at the close of Part I, Chapter III.]

CHAPTER III

STATISTICS IN THE SERVICE OF SCIENCE

Essentials of scientific method. The preceding chapter attempted to demonstrate by illustrations how perfectly good data can be used to produce entirely unwarranted conclusions. The purpose was to place emphasis upon method. The materials with which various sciences are concerned differ widely, but it is the method of dealing with facts, not the facts themselves, which makes a science. It is here that all science finds a common ground. Therefore, scientific method is fundamental and must be employed by trained minds.

Science is knowledge gained and verified by exact observation. It is the object of science to arrive at inferences and conclusions through the following procedure: (1) the careful collection and classification of facts, (2) an examination of the mutual relationships of groups of facts, and (3) an understanding of the significance of these relationships.[1] An exact knowledge of laws and causes is obtained by relating facts. It is the necessary requirement of science constantly to resubmit premises and conclusions to the test of new observations.

The impersonal character of scientific conclusions. It is difficult to make observations and to form judgments which are free from personal prepossessions and bias, which are based upon objective and not subjective considerations, which will appeal to other minds as true when the same evidence is presented. The scientist strives to eliminate by repeated observation and experiment the color which his own personality lends to the facts, and to present an argument from the facts which is as convincing to other minds as to his own. Moreover, this attitude of mind is proof against the mere eloquence of the special advocate, the plausible arguments of the propagandist, or the fervent appeals to emotion and prejudice.

The testing of premises. Premises may be assumed, without subjecting them to the test of the facts of experience, and a faultless logic may elaborate a system of philosophy which does not meet the tests of experience. For example, the type of moral philosophy which placed the responsibility for wrongdoing wholly upon the individual, to the exclusion of environmental factors, assumed that each individual is free to

[1] For a more extensive presentation of this point of view see references at the close of this chapter.

choose between right and wrong in a given situation, and did not allow for increasing or decreasing difficulty of choosing the right. If these assumptions are true, then, the conclusion as to individual moral responsibility logically follows. But it was Quetelet, the Belgian astronomer, who, in the middle of the last century, observed that types of crime were related to the surroundings of men. The facts seemed to indicate that under the same living and working conditions and social environment crimes of certain kinds occurred with astonishing regularity. A change in conditions was accompanied by a variation in the number and kinds of offenses. Crimes against property increased in times of business depression when employment decreased. Man's conduct seemed in a degree determined by his environment. When tested by the facts of experience the theory of the unconditioned freedom of the individual to choose his course of action breaks down. The premise upon which the individualistic explanation of wrongdoing rests has been modified. In the complex life of modern society it is recognized that many forces play upon the individual to mould his character and conduct. There are limitations under which freedom of choice operates on the part of any individual in a given situation, conditions imposed by the environment which are not within the control of the individual; for instance, a business depression causing unemployment so severe as to tempt an individual to steal in order to save his family from starvation. Any complete and scientific explanation of crime, therefore, must recognize a social responsibility in addition to the individual. This explanation will meet the test of experience. Such is always the requirement of scientific method.

The scientific explanation of the spread of yellow fever. Only a few decades ago, when an epidemic of yellow fever threatened a community, the local moralist explained the scourge as a punishment for the sins of the people. This theological explanation appealed to fear and not to reason.

It remained for the scientifically trained sanitarian to discover the cause of the spread of this dreaded disease. Following the Spanish-American War an effort was made to stamp out yellow fever in Cuba. A general sanitary campaign was undertaken in Havana. The usual public routes for transmission of germs — water, milk, food — were tested and supervised. Men even slept in the beds just vacated by fever patients without acquiring the disease. Still yellow fever continued to spread. It was suggested that the fatal carrier of the germ from person to person was the mosquito. Experiments were begun to test thoroughly this hypothesis. Patients having the disease were screened in order to prevent the mosquito from acquiring the germ by contact. Men risked their lives by allowing mosquitoes which had been in contact with fever patients to bite them. They were stricken and some gave up their lives in the interest of the solution of the problem of the transmission of the disease. Finally, the conclusion was reached that a certain type of mosquito was invariably associated with the spread of the yellow fever germ, as a carrier from person to person. This scientific discovery was reached by collecting and comparing the facts from many patients, and *by a process of exclusion of one explanation after another until the mosquito hypothesis alone remained unchallenged as a complete account of the transmission of the germ in each and every case.* Where the mosquito was not present as carrier, the disease did not spread. In this manner mystery yields to the training of the scientific worker in one field after another.

Induction and deduction. Having gathered and classified facts and having compared them, we may reach a judgment concerning a whole class based upon observations of individual cases, provided always the cases actually observed are representative of the larger group concerning which the general statement is made. This statement is called a generalization or an inductive inference. Such was the conclusion that the yellow fever germ is transmitted solely by means of a particular species of mosquito as carrier. Of the same character is the inference that the individual is not wholly free to choose his course of action, being limited by conditions of the environment beyond his control — and it is arrived at by gathering and comparing the facts of experience. Both of these judgments are to be distinguished from assumptions because they are based upon the facts of experience and observation.

Inductions of this sort must be verified. A given hypothesis seems to be in accord with the facts. But, do other hypotheses also explain these facts? Every effort must be made to find whether there are facts which are inconsistent with a specific hypothesis, for instance, an epidemic of yellow fever in an environment where transmission by the mosquito is impossible. Our observation may be incomplete and unobserved exceptions may occur. But verification does not rest on mere counting of cases. An inductive inference is established by eliminating other inferences. The accepted inference accords with the facts and no alternative does. In the observed cases of the spread of yellow fever there was only one circumstance common to all, that is the presence of the mosquito.

A generalization once established by inductive methods becomes the basis for deductive reasoning. If the will of the individual is not entirely free, it follows that there is a responsibility for individual conduct and morals outside of himself — a group responsibility. In turn, when this deduction is submitted to the test of experience it agrees with observed facts. The mind, by rational processes, now proceeds to make various applications of this principle of social responsibility for human conduct. The juvenile offender is not treated as a hardened criminal. He is placed on probation under suspended sentence. The probation officer looks into the environment of the offender for possible causes in explanation of his conduct. By good fortune the influence which impelled the offender to do wrong may be modified or removed.

When it has been established that a certain type of mosquito is the cause of yellow fever becoming epidemic, a relentless campaign of destruction must be waged against it. More important than any other sanitary measure is the careful screening of the yellow fever patient from any possible contact with the mosquito.

We wish to emphasize that induction and deduction are intimate parts of a complete system of investigation of facts and thinking about facts. The one cannot be separated from the other without marring both. The deductive processes are really only the reverse of the inductive. We arrive at a generalization from particular cases observed, analyzed and compared; we start with an established generalization as our hypothesis and reason to the particular applications which must agree with the facts of experience if the original hypothesis is true — the former is induction, the latter is deduction. Both processes go forward together in scientific work.

Statistics, a method of study. To most persons statistics are the masses of recorded measurements or countings — the columns of figures, the dry bones. This meaning of the term must be distinguished at once from that which characterizes statistics as the body of methods and principles which governs the collection, analysis, comparison, presentation and interpretation of numerical data. The former refers to the raw materials with which the statistician works, the latter to the scientific methods of handling these materials. The one refers to the fact basis upon which conclusions rest, the other guards at every step the accuracy of the procedures by which the dry facts are clothed with interest and meaning, and by which sound conclusions are reached. In the latter sense statistics becomes a powerful tool of science.

Descriptive and scientific statistics. Statistical data may be employed simply to describe in exact terms, or they may be so arranged as to show relations, to establish laws or to verify theories. In the first use, figures take the place of descriptive adjectives — instead of characterizing the corn crop as a bumper, the actual estimated production is stated as three billions of bushels, and the relation of this figure to the average production for the last five years is stated. *However, our chief interest centers in the use of numerical data for something more than mere precise description.* Our purpose is to make some addition to scientific knowledge, to promote a clearer understanding of social and economic organization, to explain some event in terms of related events, or to point the way to more intelligent social policies.

For example, births and deaths may be recorded merely to measure the net increase or decrease in population or to furnish legal evidence of age. But the same data may be utilized to determine the probable length of life and to construct life tables which become the basis for life insurance, or to show the relation between deaths from specific causes and the hazards of particular employments. Wages and earnings may be recorded as cost items in the business, or they may be related to data on prices over a period of time to show changes in the cost of living and the effect of these changes upon the standard of living of the worker. Data as to the mentally defective and insane may be used to determine what facilities are required in the community for their care, or a careful study of mental defects and defectives may be made with the purpose of determining to what extent such defects are hereditary. In these illustrations there may be distinguished a scientific use of statistical data to show regularities and relationships among phenomena, in contrast with the utilization merely for descriptive or administrative purposes.

The case method. When a family applies to a relief society for assistance in time of emergency that family is regarded as a case for investigation. The up-to-date relief agency wishes to know every detail about the case which may throw light upon the need for assistance, the amount required, and the possibility of making the family self-supporting again. A record is made as to the number of dependents and their ages, the earnings of each member of the family, whether the chief breadwinner is ill or has suffered from accident, the kind of dwelling and the rental paid, the habits and intelligence of the various members, and similar items. In short, a special study is made of conditions peculiar to this family in order that the necessary assistance may be furnished. The next family that applies will be found to be in a very different situation and a special study must be made in order that its needs may be met.

The worker injured in the factory applies for compensation under the law of the State. The administrative authority must determine the nature and extent of the injury in this particular case and all other facts about the accident and the injured person which may be necessary to establish whether compensation is due and, if so, how much.

The mother brings her infant to the welfare station. The child is weighed and examined by the nurse and the doctor. The mother is advised as to the special care to be given her child.

The point of view of the case method is exemplified in these illustrations. To secure a precise and comprehensive description of the case for purposes of specific treatment, a great many quantitative items are recorded. The usual attitude of mind of the case worker is not one of generalization, but of particularization. Analysis and comparison are employed to some extent, but judgments concerning the case in hand are not based on the evidence from masses of cases considered together, but rather upon comparison with other similar cases coming within the experience of the worker. Because the identity of the case is always vividly before the mind, it is difficult, if not impossible or undesirable, to eliminate the personal and subjective considerations and to base judgments upon the bare facts.

The statistical method. In contrast, the statistical method of investigation emphasizes not the characteristics *peculiar* to the individual but those *common* to many cases of a group and capable of quantitative treatment. For example, it is desired to know whether the child just weighed at the welfare station is subnormal. What is the normal weight of a baby ten weeks old? To determine this fact many healthy babies of this age must be weighed and their average weight computed. Then we can return to the weight of the individual child and answer the question.

In the case of the injured factory worker the accident occurred during his second day's experience with the particular mechanical device causing the injury. This isolated fact does not prove that inexperience or lack of proper instruction was the cause of the accident. Nor does it enable us to generalize as to the relation of experience to the occurrence of accidents. But a careful statistical investigation of a large number of accidents, classified according to the length of time the person injured had worked at the particular machine or process, may show that the rate of accident occurrence in a particular process varies inversely with the length of experience of the worker. This conclusion, once established, emphasizes the importance of better methods of introducing new workers into dangerous mechanical operations.

State legislatures in some of our commonwealths have enacted minimum-wage laws requiring the employer to pay a wage sufficient to maintain a decent standard of living. Power is granted a commission to determine the facts concerning any particular trade. This commission cannot fix compensation for each wage-earner, but it can determine a wage minimum to apply to an entire group. But how much income is needed to support an adult worker or a family at the standard of living contemplated by the law? There are two questions of fact to be determined — what items enter into such a standard budget; how much do they cost? To answer these questions satisfactorily it is necessary to collect actual typical budgets, made up of specific items and the amounts of each for which wages are spent by individuals and families in the group under investigation. From these facts an agreement may be reached as to what items must be included in order to insure health and continuous efficiency and independence to the workers. Average amounts of these items consumed in a given period may be computed from the actual budgets of many representative individuals and families. It is possible to compute the total cost of this standard budget at prevailing retail prices. The commission must then determine for the specific trade whether the employers are paying adequate wages.

These examples of statistical investigations have certain common characteristics which are in contrast to those of the case method. In each instance quantitative data are collected covering specific characteristics or conditions common to a large number of cases, as weights of infants, industrial accidents, and items of family budgets. This information is objective and impersonal. The identity of the case is important in classifying it in one group or another, in deciding whether it is typical of the group in which it is placed, and until the accuracy of the numerical data concerning the case has been verified. Thereafter the identity of the case is submerged in the group. We combine the data for many cases to arrive at a typical weight for normal infants at ten months of age; to compute an accident rate per thousand exposed workers having a given length of experience, as one week, one month or one year, and working at a specified process in the factory; to discover the average requirements for a family of five persons in order that its members may maintain a decent standard of living. In short, from the detailed numerical data representing many cases we generalize in terms of averages or types for an entire group. *The attention is fixed upon the forest and not primarily upon the trees. The interest of the investigator is concentrated upon the typical, not upon the unusual cases.* The individual cases form the basis of our generalization, but we must judge the individual in the light of group behavior, for example the weight of the infant compared with the normal weight for children of that age.

The non-mathematical aspects of statistics. The principles of statistical method are founded in mathematics. To devise, test, and apply refined methods requires a knowledge of higher mathematics. *It has been remarked that nine tenths of statistics is common-sense.* Whether common-sense makes up nine tenths or five tenths of statistics does not concern us here. Certain it is that every scientific worker and, above all others, he who deals with quantitative data, must apply all possible checks of consistency to his original data and to his results. Because facts and results are stated in precise numerical form they make an unduly impressive appeal to the uncritical. *A healthy skepticism is probably the most essential quality of a statistician.* For example, in a laboratory exercise a student calculated the average age of a group of children whose ages were about equally distributed above and below eight years and ranged from five to fourteen years. By two recognized methods of computing an average he obtained the widely different results of eight and one half and fifteen and one half years. Both values were turned in without comment as equally credible. Even casual inspection of the original data as given should have enabled the student to discard the fifteen and one half years as an impossible result. However, his attention was so concentrated on the use of a mathematical process and the obtaining of a precise numerical result that he neglected entirely the checks of common-sense and consistency. Indeed, this is all too likely to happen when the worker is using formulæ and figures.

Furthermore, when a formula is applied and all the calculations are accurate it does not follow that the resulting figures express the truth. For example, the infant death-rate is usually expressed as a certain number of deaths under one year of age per thousand born during the particular year. The formula for calculation is:

$$\frac{\text{Deaths under one year of age for given year} \times 1000}{\text{Recorded births for the same period}} = \left\{\begin{array}{l}\text{Death-rate of infants}\\ \text{for the given year}\end{array}\right.$$

But in many communities not all births are reported. Perhaps only eighty out of every hundred actual births are recorded. In this case the original data below the line in the formula are inaccurate to the extent of twenty per cent. In spite of this situation the infant death-rate is computed, sometimes to one or two decimal places, and the result is published as if it were the true rate. Furthermore, it is compared with the rates of other communities in which births are recorded with various degrees of accuracy. The fact that the computation is correct and that the result is stated in one or two decimals gives the consumer of the statistics a false sense of accuracy. Mathematical methods may be used to measure the degree of accuracy of the original data, but sometimes the mathematical procedure itself covers up these inaccuracies.

The problems most difficult for one working with quantitative data are those concerned with the decisions as to whether particular methods and procedures apply to the available facts, and what conclusions are justified from the data.[1] *Good judgment, broad knowledge and experience, and common-sense are the most valued possessions of the statistician and research worker.* The more complete the mathematical training associated with these qualities, the better equipped will be the statistician.

STATISTICS IN THE SERVICE OF THE SOCIAL SCIENCES

Statistical method bears a relation to the social sciences somewhat analogous to that between experimental or laboratory method and the natural sciences. Social statistics had their beginnings in the effort to adapt to the study of society and social relations methods similar to those found to be effective in the study of nature.

The experimental method. When the layman thinks of science, he is apt to have in mind the experimentation in physics, chemistry, or bacteriology by which a better understanding of Nature and her forces has been slowly evolving through the method of trial and error. Experimental research in chemical, engineering and electrical laboratories has resulted in inventions which have revolutionized modern industrial life.

It has been characteristic of the experimental method to devise laboratory apparatus for exact observation, measurement, recording and counting of data. These are exemplified in the microscope, the hair balance for minutely accurate weighing, the delicate instruments for recording earthquake shocks at great distances, the moving picture, and the calculating and tabulating machines. No less fundamental has been the working out of methods for keeping all variable factors in the given experiment under control, in order to observe the changes which occur in one variable while changes of a known character take place in a second variable, all other factors affecting the experiment being kept constant. This is the usual laboratory procedure for studying the relationship between variables.

Limitations on the experimental method in the social sciences. In the first place, the units with which the social scientist is concerned are not alike and cannot be managed for observation and experiment with the same certainty and ease as can chemicals in a test tube over a gas flame. Human nature makes the control of many of the variables related to welfare difficult, if not impossible.

A second difficulty arises from the fact that many related factors enter into a given social situation making it complex. Since, as has been re-

[1] J. M. Keynes, in his *Treatise on Probability*, published in 1922, warns the statistical worker as follows (p. 382): "There is no more common error than to assume that because prolonged and accurate mathematical calculations have been made, the application of the result to some fact of nature is absolutely certain."

marked, these variables are difficult to control, it may happen that some uncontrolled variable is responsible for the observed result. For example, it is decided to establish infant welfare stations in certain sections of a city to furnish a better milk supply, to provide for an examination by a nurse and a physician, and to instruct the mother in the care of her child. The infant death-rate is lowered to a marked degree and the welfare station is given the credit for this noteworthy achievement. But other conditions have not remained the same while welfare stations were being developed. General sanitation has been greatly improved in the community. Wages have increased and the standard of living has been raised. Health education has been extended more widely, sponsored by various agencies, official and private, social and business. In contrast with the experimental laboratory, it has not been possible to keep all other conditions constant except the one under investigation. To what, then, is the lowered infant mortality really due? Many a logical fallacy enters into our conclusions through failure to weigh the influence of the uncontrolled variables in a given situation.

Some one hastens to attribute the inefficiencies and difficulties of railway management in the United States during and following the World War to Government control. But Government control never had a fair opportunity to show what it really could accomplish, because other conditions did not remain the same as before. In fact, financial and equipment conditions of the railways were rapidly approaching a crisis before the Government entered the situation. Under war pressure and conditions of scarcity of labor and capital it was too much to expect the Government to remedy the evils already present. Besides, many obstacles were placed in the way of efficient Government operation by those interested in seeing the experiment fail. Many and varied were the factors which could not be controlled in this situation. Whatever else the venture may have been, it was not a fair test of the effectiveness of Government ownership.

There is still another difficulty in undertaking experiments in human welfare. A generation may be required to produce results. Deliberate changes are wrought so slowly that in the meantime many social and economic factors and relationships, other than those under control, have been changed concurrently. To what factor or factors, therefore, can any resulting improvement be fairly attributed?

Notwithstanding the difficulties just discussed, the experimental method can be employed with success in some situations, provided due caution is exercised. For instance, some employers are beginning to take the experimental attitude toward their labor force, as they have done from the beginning with their mechanical equipment. Rest periods are introduced at mid-morning and mid-afternoon, and the effect upon hourly and total daily output is observed. Frequently the results reveal a larger total output and better quality of work with less drain upon the energies of the worker. In certain processes the evidence from these experiments indicates a decrease in the frequency of industrial accidents, due to lessened fatigue and better muscular coördination. These results mark the beginning of a new science of industrial physiology. Thus, the argument for the regulation of the length of the working day and for the provision of more humane working conditions finds strong support in the self-interest of the employers as well as in the welfare of the workers.

Dependence of the social sciences upon statistical method. Since the experimental method in the social sciences has the serious limitations already explained, some method must be employed which offers the precision, the objectivity and the possibilities of comparison and generalization which have been so productive in the laboratory sciences. Statistical method meets these requirements. The relationships between the significant factors in human associations may be studied by means of the comparison of groups of data which can be classified and measured, but cannot be controlled easily for experimental purposes.

In a very real sense the modern community furnishes a laboratory where action and reaction are exhibited in endless variety. The scientific observer finds at hand many nationalities and races living and working side by side; a multiplicity of employments pursued under differing degrees of healthfulness, continuity and efficiency; a wide range of living conditions, earnings and standards of living. The investigator enumerates or measures the characteristics of social, economic or political life which are capable of quantitative treatment, as population, births, deaths, accidents in industry, trade, prices, wages, employment, family budgets, production, waste, profits, crimes, and votes — a bewildering array of facts. It is his business so to classify and arrange these facts that their relations may be studied in the hope that knowledge will be increased and welfare promoted.

The statistical method becomes an instrument of discovery like the microscope, revealing relations and possible causes heretofore unobserved. The individual measurements are subject to such marked variations that methods for dealing with masses of data must be employed to show a definite and uniform tendency for variations in one condition or characteristic to depend upon or be associated with changes in other conditions. For example, the examination of individual cases may reveal no uniform relation between the number of hours worked by the injured and the occurrence of industrial accidents. However, when a large number are classified according to the number of hours worked before the injury and the kind of work done, it may become evident that in a particular employment there exists a tendency for accidents to increase in frequency with the increase in the length of the work period. It is only by the application of statistical methods that we can hope to show the uniformities and relations in the infinite variety of the facts of social life, and on the basis of this knowledge to formulate empirical laws. The complexity of relations requires the investigator to keep in constant and close touch with concrete facts.

THE SERVICE OF STATISTICS TO ECONOMICS AND SOCIOLOGY

1. *Quantitative measurements furnish a precise and objective description of economic and social organization and of the changes which occur from period to period.* Mere qualitative description is not enough to constitute a science in the sense discussed earlier in this chapter. Description may convey a general idea of the banking organization of the community but the operations of the system are revealed by quantitative statements as to the amount of loans and discounts; the character and amount of reserves and their relations to loans and discounts, the volume of notes issued and the nature and amount of their security. By these and similar facts the strength or weakness of the banking system is measured.

Unemployment and labor turnover may be described in general terms as prevailing evils in our industrial organization, but a working knowledge of these problems depends upon exact quantitative information. How much unemployment exists in different trades and industries, at different seasons and under changing conditions of business prosperity and depression? What is the proportion of part-time employment, the duration of unemployment and how do these affect the annual earnings of the workers? What are the types of unemployed and the relative numbers in these various groups? When many are out of work in one trade or in one geographic section, what is the situation in other employments and in other sections? Why do workers leave their jobs and what is the relative importance of the various reasons when classified and compared? These facts and classifications are fundamental in judging the nature and burden of unemployment and in formulating policies of control.

It is not enough to observe that foreign populations have been flowing into the United States in increasing numbers in the quest for greater economic opportunity. From what countries and in what relative proportions do they come? Where do they settle and what do they do to earn a living? What proportion fails to maintain independent standards of living? Is their residence permanent or temporary? What proportions among the various nationalities become citizens and how do they vote? What is their birth-rate and size of family as compared with the native population? These facts are necessary in deciding upon a wise immigration policy and in planning an Americanization program.

It is not sufficient to know that sanitation and the conservation of life has become an American ideal. The modern science of sanitation depends upon statistics of sickness and mortality. What causes of illness and death are decreasing and under what conditions? What causes show an increase, and why? What are the hereditary and environmental conditions dangerous to health?

2. *Economic and social organization and activity involve complex relations of group phenomena.* Therefore, comparison is necessary in order to comprehend the significance of these relationships in a given situation. As a rule group phenomena can be compared only by statistical methods. To illustrate, in a certain locality it is desired to ascertain whether the workers are as well paid in one industry as in another requiring like skill. It will be inadequate to make inquiries of a few individual wage earners in the two industries because the kinds of work done and the wages paid vary. The wages of all the workers in each industry or of a representative sample must be compared. This necessitates a statistical inquiry and a judgment from mass data properly classified and summarized.

Statistical methods are useful in eliminating or standardizing some of the complex factors in social causation while others are being compared. Seattle has a much lower general death-rate than Richmond, but this does not accurately reflect the difference in healthfulness between the two cities. In the former city are much smaller proportions of children under five years and old persons over sixty years among whom the mortality is always high. Furthermore, Richmond has a larger percentage of colored in its total population and the death-rate of the colored population is usually much higher than that of the white. These differences in age and race always affect the mortality rates and appropriate statistical methods must be employed to render the rates of the two cities comparable.

3. *Society is not static, but kinetic.* Progress involves change. Relations once established may not remain the same. Statistical data and the methods of analysis are essential in measuring changes from period to period. Some of the most significant relationships are between the movements of series of facts over periods of time. The record of economic and social facts should be continuous.

Malthus in his day observed that population tends to increase faster than production of food and raw materials. As a result he stated the empirical law that population tends to increase in geometrical progression while food production increases only in arithmetical progression. It follows logically that the income of the common laborer is fixed at the level of subsistence. If wages rise above this level, population increases until they fall again, and if they fall below this level starvation results. Such a generalization needs constant checking against the facts of experience. When Malthus first formulated this law he did not take into consideration voluntary control over the birth-rate. To-day this voluntary control over the size of the family has become the most effective check on the growth of population. The birth-rate almost everywhere shows a marked tendency to decline, while wealth and productive power have increased.

How do the *real wages* of workmen change from period to period, as measured in goods which their money income will buy? The answer requires statistical data on money wages for the period in question and on retail prices of goods and services entering into the ordinary family budget. The comparison from year to year of the changes in these series of data will show whether real wages are increasing or decreasing.

The business man is especially interested in series of statistical data which indicate the movement of prosperity and depression and which enable him to forecast the probable events of the immediate future. By comparing the internal facts of his own business with the general movement in the community or the country he is able to arrive at wise decisions and so to avoid losses. If he waits until his own declining sales warn him of impending depression, it is usually too late.

4. *Quantitative data and statistical methods of analysis and comparison aid in demonstrating the relativity of social and economic generalizations.* A relation or explanation which holds true to-day may be greatly modified as time passes. The explanations of the determination of wages and profits and of the fixing of prices under the conditions of free competition are no longer true when monopoly has entered or when government regulations are put in operation. Hypotheses and generalizations are modified or abandoned if they do not meet the test of new facts.

The law of population referred to must be restated in the light of recent experience. Malthus could not foresee, when he first stated the law, the rapid development of inventions and the applications of steam power to machinery and to ocean transportation which made England a manufacturing nation and enabled her to bring food and raw materials from many lands in exchange for manufactured goods. Population continued to grow rapidly but production and wealth increased still more rapidly, and a rising standard of living became possible. As time passed the contrast was no longer between population and the food supply, but between population and efficiency in production and trade. Besides, with this rapid growth of capital and wealth has been associated the voluntary check upon the size of the family. The emphasis has been changed to the achievement of a rising standard of living. The pressure of population is still a reality but is exerted not against the barrier of the level of subsistence, but against the barrier of a standard of living, which for the common labòrer is well above subsistence.

5. *Statistical data and their analysis furnish the bases for wise social action, for restrictive legislation, and for better administration.* Modern preventive activity and movements for social and economic reform should be founded upon the understanding of causes. It is the goal of statistical method to make clearer the relations of cause and effect and to measure results. The accumulation of exact information, much of it statistical, concerning the exploitation of the members of industrial society and the wastes of the individualistic policy has led to the interference of government in industrial relations. In confirmation of these statements one needs only to point to the development of accident prevention, to the general public health movement, and to the measures for the protection of women and children in industry.

STATISTICS IN THE SERVICE OF EDUCATION

Educational experts, utilizing statistical methods, have made highly important contributions to the science of human nature and behavior. They have demonstrated that individuals differ widely in respect to traits of intellect and character. The causes of individual differences are suggested by relating them to the influences of sex, race, family, education, and environment. The point of view is inductive or experimental and the method is statistical. It follows that a system of education which treats persons as if they were alike either fails utterly to reach the needs of a large number or tends to level out the variations upon the continuance and development of which progress depends.

The schools of former generations gave attention to a selected group. Recently, compulsory attendance has brought all types of children into the schools. A fuller curriculum, a longer school year and higher standards of work have brought to light the maladjustments caused by the older uniform methods. In the great industrial centers the school population is drawn from widely different economic classes with varying environments which register their effects upon the physical and mental life of the children. It is observed that many pupils leave school at an early age and many more are retarded in their school progress.

Physical examination of pupils has revealed many defects. Exact information concerning the physical and mental characteristics of each school child is destined to play an important part in shaping educational policy, in relating the work of the school to the needs of the particular child, in seeking to make the school an agency for correcting physical defects, and in bringing about the removal of the causes of these defects found in the environment of the child outside the school. A continuous, quantitative record of both physical and mental conditions of the school child is necessary. These records, when analyzed and compared, show a relation between defects and school progress; between the condition of the child as we find him in the school and the surroundings from which he has come. In this manner a scientific attempt is made to improve the educational opportunities of individual children.

SUMMARY

Instead of attempting a formal definition of statistics at this point in our study, let us summarize the nature and utility of statistical method. Statistics is a method of investigation which involves (1) exact measurements or quantitative estimates, (2) careful recording of results in classified form, (3) analytic scrutiny and treatment for purposes of comparison, and (4) judgment of the evidence and generalization from it where possible.

Statistical method involves the essentials of general scientific procedure, emphasizing the attitude of mind which insists on basing conclusions upon facts accurately collected and properly classified, and on understanding the significance of relationships between groups of facts.

Statistics deals with mass data. Therefore, not all numerical statements are statistical. The items must be so chosen and the number must be large enough to be representative of the group under investigation. Statistical method aims at discovering and presenting uniformities or types which become the bases of general statements concerning phenomena. These generalizations from numerical facts take the place of assumptions and traditional beliefs and customs. Analysis of mass facts creates a foundation upon which to build industrial, educational, political and social policies, and furnishes a testing instrument of results.

Quetelet is sometimes called the father of modern statistics, because his work did much to turn attention from mere description to the use of quantitative data for purposes of explanation, for measuring similarities and differences, for the study of relationships, for defining antecedent and consequent, cause and effect. To-day the expert in business and the trained student and research worker in economics and sociology must know something of statistical method and technique. Knowledge of the methods of collecting and treating quantitative data and some practice in their application emphasizes the scientific point of view in approaching any problem; cultivates the critical faculties; and develops ability to sift evidence on the basis of the true and the false, the adequate and the inadequate, rather than from the viewpoint of conformity to predetermined ideas, beliefs and traditions.

READINGS

(These readings are suggested for the purpose of presenting different points of view from which writers have approached the study of statistical methods.)

King, W. I., *Elements of Statistical Method*, part I.
Mills, F. C., *Statistical Methods, as Applied to Economics and Business*, chap. I.
Pearson, Karl, *The Grammar of Science, Physical*, part I, 3d ed., chap. I.
Yule, G. U., *An Introduction to the Theory of Statistics*, 6th ed., "Introduction."
Bowley, A. L., *Elements of Statistics*, 4th ed., part I, chap. I.
—— ——, *An Elementary Manual of Statistics*, part I, chap. I.
Secrist, Horace, *An Introduction to Statistical Methods*, chap. I.
—— ——, *Readings and Problems in Statistical Methods*, chap. I.
Jerome, Harry, *Statistical Method*, chap. I.
Rugg, H. O., *Statistical Methods Applied to Education*, chap. I.
Mayo-Smith, Richmond, *Statistics and Sociology*, chaps. 1, 2, and 3.
Giddings, Franklin H., *The Scientific Study of Human Society*, chap. 12.
Whipple, G. C., *Vital Statistics*, 2d ed., chap. I.
Pearl, Raymond, *Medical Biometry and Statistics*, chap. I.
Jones, D. C., *A First Course in Statistics*, chap. 1, p. 4.

REFERENCES

Columbia Associates in Philosophy, *An Introduction to Reflective Thinking.* (Assists the student to relate closely inductive and deductive processes.)

Moore, Henry L., *Laws of Wages*, "Introduction." (The scientific approach to economic theory.)

Mills, F. C., "On Measurement in Economics," from *The Trend of Economics* (R. G. Tugwell, Editor), chap. 2.

West, Carl J., " Value to Economics of Formal Statistical Methods," *Quarterly Publication of American Statistical Association*, September, 1915.

Quarterly Publication of American Statistical Association, March, 1914. (Articles on the service of Statistics to Sociology, to Economics, to Biology, and to History, each by an authority in the field.)

Giddings, F. H., "The Measurement of Social Forces," *The Journal of Social Forces*, November, 1922.

Willcox, W. F., " The Need of Social Statistics as an Aid to the Courts," *Quarterly Publication of American Statistical Association*, March, 1913.

Hart, Hornell, "Science and Sociology," *The American Journal of Sociology*, November, 1921.

Koren, John (Editor), *The History of Statistics. Their Development and Progress in Many Countries.* Published for the American Statistical Association by Macmillan, New York, 1918. (Devoted mainly to official statistics.)

Meitzen, August, *History, Theory and Technique of Statistics.* Translated by Roland P. Falkner, Annals of the American Academy of Political and Social Science, Supplement, March, 1891. (Part I deals with the history of statistics from earliest times.)

John, V., *Geschichte der Statistik.* Enke, Stuttgart, 1884. (The best history of statistics to the early nineteenth century.)

Bertillon, J., *Cours élémentaire de statistique.* Société d'éditions scientifiques, 1895. (Useful outline of the history of official statistics in different countries.)

Merz, J. T., "On the Statistical View of Nature," *A History of European Thought in the Nineteenth Century,* 2d unaltered edition, vol. II, chap. 12 (London, 1912).

Mitchell, Wesley C., "Quantitative Analysis in Economic Theory," *American Economic Review,* March, 1925. (Presidential Address at the Annual Meeting of the American Economic Association.)

[Details as to publisher and date of publication of the readings and references given at the close of each chapter are found in the alphabetical list in Appendix A.]

PART II

CLASSIFICATION AND DESCRIPTION OF MASS DATA

The problems of natural science have required the invention of a calculus of mass phenomena that will probably yield its best results when applied to the material of the social sciences. The wealth of the statistical material . . . is itself a source of embarrassment. To utilize it for scientific purposes, it must be described in brief, summary formulæ, and these formulæ must be arranged upon a plan of increasing complexity so that it will be possible to pass from accurate descriptions of mass aggregates to the relations between the aggregates themselves.

Henry Ludwell Moore

Laws of Wages. The Macmillan Company
New York, 1911, pp. 4 and 5.

CHAPTER IV

CLASSIFICATION OF STATISTICAL DATA

First steps in statistical analysis. To identify and describe a subject for investigation means "to resolve it into components or elements of place, time, circumstance, quality, magnitude, activity, behavior or function, coexistence or sequence." [1] Numerical data, to be useful for scientific purposes, must be arranged in classified form. The bases of these groupings are likeness and difference.[2] Common characteristics, as height, weight, age, nationality, cause of death, and family income, are enumerated, measured, or estimated. Degrees of difference in quality and quantity determine the number and limits of the groupings. Classification is especially important in the social sciences because of the many factors which affect a given situation and because the measurements show such wide variations.

For example, in a single factory there are performed many kinds of work requiring different degrees of skill and for which widely varying wages are paid. There is a combination of hand work and machine work; some operations are paid on a time basis by the day or hour, and some at piece rates; both men and women are employed, and union and non-union workers. The investigator on wages secures his raw statistical material from the paysheets of the factory. He groups the data under categories of sex, kind of work, degree of skill, time work, piece work, and the like. In each of these groups the weekly earnings of the individual workers, when recorded, vary considerably. Therefore, the workers are classified further according to the size of their earnings. By this arrangement of data the most common wage paid in any group of workers is revealed, and how much the wage of each individual varies from this amount. It is desired to know how much and why wages vary. Do the wages of men and women differ for the same work? Does the rate of pay vary directly with the skill of the worker? What is the effect of labor organizations on wages? How does the length of experience in a given employment affect the wage paid? Does the method of wage payment affect output, and how? Is there economy in paying higher wages or in shortening the hours of work? The scientific investigator is not content

[1] Franklin H. Giddings: "Societal Variables," *The Journal of Social Forces*, March, 1923.

[2] Franklin H. Giddings: "The Classification of Societal Facts," *The Journal of Social Forces*, January, 1924.

to classify wages merely as a variable in amount, but is interested in relating wages to the various factors which may influence them.

The health of a city is measured by certain quantitative relations, as the number who are ill or die in a year in each thousand of the population. This fact gives very little useful knowledge about the health conditions of the community. Diseases affect different age groups in very different degrees. Therefore deaths must be classified by age periods and cause of death, if we would learn what diseases are characteristic of different periods of life. Likewise illness and mortality are unequally distributed among workers in different occupations, and classification of vital facts by occupation as well as by age groupings becomes essential if we would discover hazardous employments. Grouping mortality data according to nationality and race shows great differences. Some racial stocks appear to be more susceptible than others to specific diseases, as tuberculosis. Many other illustrations will occur to the student of social and economic conditions which indicate the fundamental character of classification in scientific analysis and in the effort to establish and to interpret relationships.

Simplification of data. We have already emphasized that statistical methods are concerned with characteristics common to masses or groups. The large number of items measured or enumerated makes necessary some sort of classification. For example, if each person in a group of five hundred is requested to write his age on a separate card and these records are passed rapidly through the hands of the observer, without any attempt at orderly arrangement, only a vague idea is obtained concerning the age distribution. However, by the simple device of sorting the cards into piles, each of which includes the ages of those within a five-year period, a more accurate idea is gained.

Likewise the weekly earnings of five thousand workers in a factory may have been accurately recorded. Observed as separate items without orderly arrangement, they reveal individual differences, but no concept is gained of a type wage from which the individual items are variants. No accurate statement can be made concerning the proportions earning more or less than a specific amount. A similar number of ungrouped wage items obtained from the records of another factory, when compared with the first, have no meaning to the observer. We can neither judge the significance of masses of data, nor compare them with other similar masses until they are grouped according to kind and size and until they are arranged in orderly sequence.

In January, 1920, the Federal Census enumerators recorded the ages of over five and a half millions in New York City, but in order to ascertain the number of children of school age in the city at that time it was necessary to classify the population by specific age groups.

The basis of valid comparison. In Chapter II illustrations were given of errors arising from the comparison of data which were not comparable. Naturally we are interested in avoiding these errors. For example, it is desired to determine the relative fertility of two groups in our population, the native-born and the foreign-born. For this purpose it is not adequate to compare the number of births per thousand of the entire group, including men, women, and children. This crude birth-rate is misleading because there is a larger proportion of married women of child-bearing age among the foreign-born. Therefore the crude rate will be relatively high for this class of the population, irrespective of influences limiting the size of families. Classification of the population by sex, age, marital condition, and nativity is required to permit the calculation of refined birth-rates, which alone are comparable, as the number of births in a year per thousand married women fifteen to fifty years of age.

Sometimes occupations are ranked in respect to their relative hazards by comparison of the number of deaths in a year per thousand exposed workers in each occupation. In a comparison between different employments no account is taken of differences in the average age of the employees. If the employees are older in one occupation than in another the death-rate among the workers of the former will tend to be higher, regardless of special hazards. On this account, for valid comparison of mortality rates in different occupations, classification of deaths according to the age of the deceased as well as the kind of work done must be made.

Thus, classification of data in statistical analysis is a method of securing such degree of homogeneity as will make the comparisons valid. It would be useless to compare the average wage of two groups of workers in one of which only men were included and in the other both men and women.

Attributes and measurable characteristics. Two fundamental principles of classification should be distinguished. (1) Groupings are made according to specific attributes.[1] The differences between cases are *qualitative*, not quantitative. Classes may be formed according to the presence or absence of a certain characteristic — a twofold division, as skilled or unskilled, sane or insane. As a rule, however, classification involves more than a twofold division. Data representing a specific attribute are grouped in manifold subdivisions, as *nationality* according to the particular country of birth, or *occupation* defined according to the kind of work performed. The numerical character of the data in these

[1] G. U. Yule: *An Introduction to the Theory of Statistics*, 6th ed., 1922. Part I discusses the methods adapted to treatment of data of this character.

classes arises *solely* from countings or enumerations of the cases belonging to each category as defined more or less arbitrarily by qualitative distinctions. Detailed statistical methods appropriate for this type of data are not presented in this treatise.

(2) Groupings are also made according to *quantitative* differences in some measurable characteristic, as age, or amount of income. *Magnitude,* not qualitative distinction, is the basis of classification. The specific characteristic exhibits a range of variation in values from a minimum to a maximum amount. The numerical data are obtained by enumerations, measurements or estimates.

In the same investigation both principles of classification may be employed. For example, workers are grouped according to qualitative distinctions in the kind of work or the degree of skill. Then the unskilled group is further classified quantitatively according to the amount of weekly earnings of each individual.

Moreover, in the social sciences it is important to devise means of measuring differences, changes and relationships for which established units of amount are not available, as they are in the case of height or weight. Effort is made to transform mere qualitative differences into quantitative by the use of appropriate numerical scales. Mental abilities and human traits, which formerly were described under qualitative categories, are now described and related with greater precision by the use of scales.[1]

Types of series in the classification of quantitative data. If the purposes of scientific investigation are to be accomplished in the best possible manner, statistical methods must be adapted to the character of the data treated. There are fundamental differences between series of quantitative data which should be recognized in applying appropriate methods. *Series is a general term* used to describe a succession of enumerations, measurements or estimates without specifying the particular form of arrangement of the data. *Types of series should be distinguished according to the fundamental relations of the items in time and space.*

1. Observations may be made from the point of view of time relations, as prices of commodities collected monthly over a period of years; quantity of output in a factory at successive hours of the day; amount of exports and imports month by month; immigration year by year. The numerical data are used to describe a succession of events located at specific intervals of time. Each value may be an average of many items, as in the case of prices; or merely countings, as in the case of annual immigration; or estimates, as in the case of business forecasts.

[1] Edward L. Thorndike: *Individuality.*

2. Data may be collected and analyzed for the purpose of describing *a cross section of phenomena at a specific time. Then space relations are of fundamental importance.* Arrangements of facts on this basis produce different kinds of series.

A. Geographic location may be the essential consideration. Each value in a *geographic series* may be an average or a ratio, as yield of wheat per acre or percentage of illiteracy in the population; or each value may be the result of an enumeration or estimate, as the population of States, the amount of capital and the number of wage-earners in the manufacturing industry by geographic areas.

B. Classification of data describing a cross-section may emphasize varying amounts of a specific characteristic, as age or income, rather than the location of the units in a particular place. This type of series is illustrated by wages taken from the paysheets of a factory; heights and weights of many school children; the ages of the foreign-born. The essential characteristic of these data is *variation in magnitude* from a lowest to a highest value, or *vice versa*. The following chapters of this book set forth methods appropriate for the classification and treatment of quantitative data of this type.

Adaptation of methods to the specific type of series. 1. Change over a period of time is characteristic of modern social and economic life. Therefore, in a time series where differences from period to period are fundamental, some statistical device must be invented to measure change. An index number meets this need. The construction and use of this device will be discussed in a later chapter.

In analyzing time series different sorts of movements must be distinguished. (*a*) Within any year there occur seasonal differences revealed by monthly figures, as the numbers employed month by month in the garment trades, or the rise in the number of infant deaths during midsummer. Statistical analysis must determine whether there exists for a specific type of data a recurring seasonal swing, and, if so, its amount. (*b*) When longer periods of time are considered, increases and decreases recur in more or less regular wavelike movements. In business these are described as periods of prosperity and depression — business cycles. (*c*) Over still longer periods in many time series there appears a general movement, or trend, either upward or downward, which persists in company with the wave movements of the cycle just described. For example, the bank clearings of New York City over a long period of years show a persistent rise in amount, due to the increasing use of credit and the growth of industry and trade. In contrast, the general death-rate over the same period shows a marked decline, indicating progressive control over human wastes.

Prices of commodities show all three of the movements just described, seasonal, cyclical, and the general trend upward or downward. To guard the accuracy of inferences drawn from the data, and to establish a basis for the prediction of future events, appropriate methods must be employed to analyze all these short-time and long-time changes. Figure 1 illustrates the types of variation found in a time series. The methods for handling time series will be discussed and illustrated in Chapter XIII.

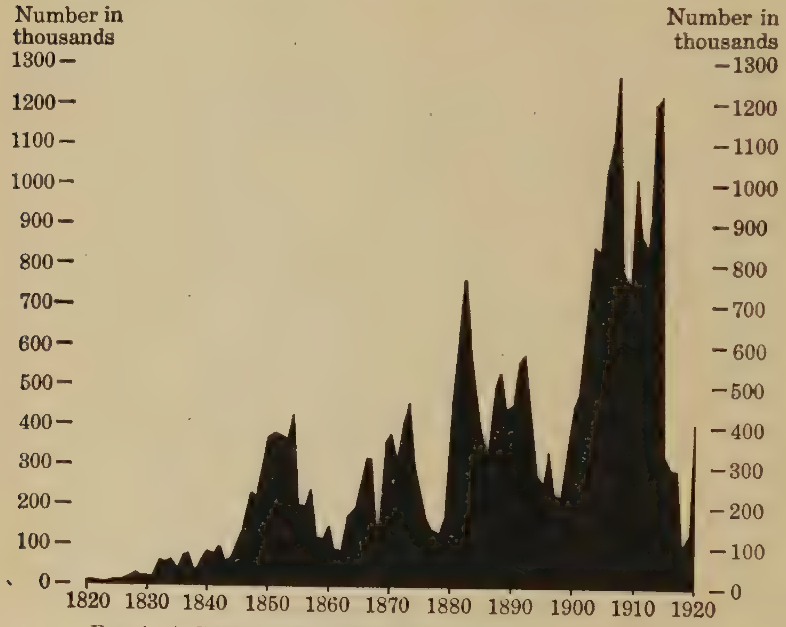
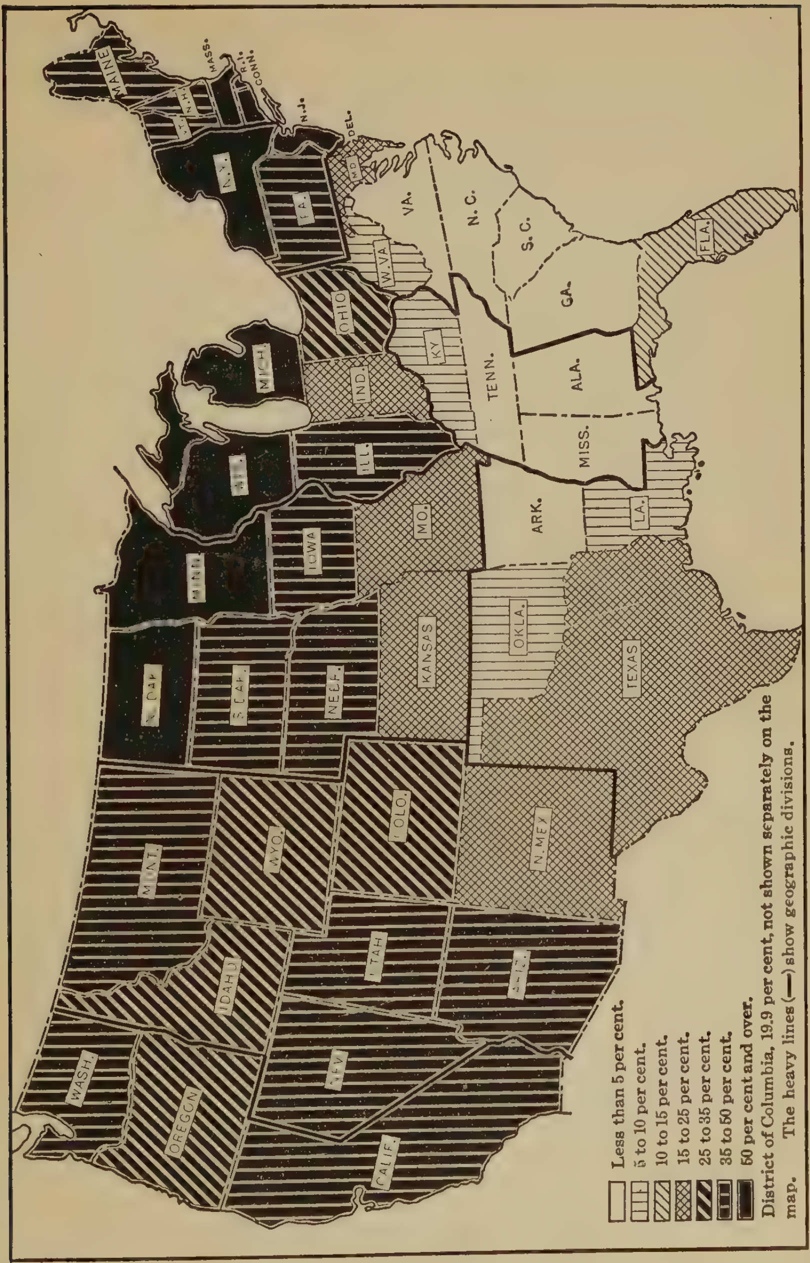


FIG. 1. A CENTURY OF IMMIGRATION TO THE UNITED STATES

2. Methods must be adapted for classifying, comparing and presenting the data when differences according to location, rather than changes in time, are made the basis of analysis.[1] Where conditions are so varied as in the United States contrasts are striking even within the same city and are strongly emphasized between the urban and rural sections. Wide geographical distribution of the population and of industry and a varied physical environment produce significant differences in activities and conditions. Under these circumstances it becomes essential to record and analyze facts in cross-section. How do wages and employment vary in

[1] A complete discussion of geographic series is not attempted in this treatise. Reference is made in particular sections of the book to the application of statistical methods tc this type of data, as in the chapters on the mode and on graphic presentation.



(*Courtesy of Bureau of the Census, Fourteenth Census, Population, Volume II, p.* 23.)

FIG. 2. PER CENT OF FOREIGN-BORN WHITE AND NATIVE WHITE OF FOREIGN OR MIXED PARENTAGE COMBINED IN TOTAL POPULATION, BY STATES, 1920

different localities? Why is the death-rate in one section of a city so much higher than in another? Why do food prices vary in different cities? Index numbers also may be used to measure these geographic differences in wages and prices. Frequently the average for the entire country is not so important as is the contrast between two or more sections. Graphic devices may be employed to excellent advantage in presenting this type of data, especially maps shaded in proper contrasts. Figure 2 illustrates the graphic representation of a geographic distribution.

Caution in comparing data for different areas is necessary if we are to avoid wrong inferences. (*a*) Sometimes confusion results because of different definitions of the unit which is counted, measured or estimated. For example, different States do not define an industrial accident, for purposes of record, in the same terms. Numerical totals of accidents do not mean what they appear to mean because in one State all accidents are recorded and in another only those causing a specified loss of time to the injured. Likewise, the number of arrests for certain offenses in one community when compared with the number for like offenses in a second community do not indicate, necessarily, the relative moral conditions, because the one community may enforce the law more strictly than the other. (*b*) Sometimes the difficulty lies in the character of the groups compared. The comparative healthfulness of two cities is often measured by the general death-rate per thousand of the entire population in each. The rate for one city may be relatively high merely because it contains a larger proportion of children and old persons. The same source of error is found in general comparisons between urban and rural populations. (*c*) Sometimes units measured or enumerated in different localities, although having the same name do not really mean the same thing. In recording the retail prices of commodities such as flour, butter, and eggs in the same city or in localities widely separated, it is not easy to be sure that the same grade or quality is being quoted. It is simpler to collect wholesale prices because the markets are not so varied and many commodities, as wheat and cotton, have been carefully graded according to a generally accepted scale.

State lines in the United States prove to be a serious obstacle in the development of comparable statistics. Each State and locality has been to a great extent independent in the classification and presentation of essential social and economic data. The need is for agreement and for coöperation between statistical organizations in the several States on fundamental facts to be collected, on methods of gathering and analysis, and on the forms for presentation. The problem is largely one of uniform classification.

When variation in the amount of the characteristic selected for measurement is the essential consideration, the arrangement of the data takes the form of a *frequency distribution.* The following chapters are devoted to a discussion of methods of forming these distributions from the original data and methods of describing them by means of averages, measures of variability, and graphic representations.

SUMMARY

The purpose of investigation is to describe and to relate groups of data. Statistics measure mass phenomena and the individual items must be arranged in proper groupings and relations, according to their similarities or differences. Classes may be determined according to qualitative or quantitative distinctions. Series of quantitative measurements may be arranged to show variations over a period of time, or on the basis of geographic differences, or according to variations in magnitude. More or less arbitrary distinctions and limits must be determined, always keeping in mind the practical utility of groupings for purposes of summarization, comparison and interpretation. Facts characteristic of a given time or place must be related to those characteristic of other times and places. Therefore, the continuity of the record and the comparability of recorded data are essential. Finally, appropriate statistical methods must be adapted to the different types of statistical data.

READINGS

Jones, A. L., *Logic Inductive and Deductive, An Introduction to Scientific Method,* chap. 3.

Bowley, A. L., *The Measurement of Social Phenomena,* "Introduction," and chaps. 4 and 5.

Giddings, F. H., *The Scientific Study of Human Society,* chap. 4.

Day, E. E., "Classification of Statistical Series," *Quarterly Publication of American Statistical Association,* December, 1919.

Jerome, Harry, *Statistical Method,* chap. 3. (Analysis and classification are preliminary to gathering and tabulating data.)

Rugg, H. O., *Statistical Methods Applied to Education,* chap. 4.

Zizek, Franz, *Statistical Averages.* Translated by Warren M. Persons, part I, chap. I.

REFERENCES

Yule, G. U., *An Introduction to the Theory of Statistics,* 6th ed., part I. (Classification on the basis of quantitative differences and qualitative differences.)

Whipple, G. C., *Vital Statistics,* 2d ed., chap. 9. (Classification of Occupations and Causes of Death. Occupations are classified also in Fourteenth Census, *Population,* vol. IV.)

Pearl, Raymond, *Medical Biometry and Statistics,* chaps. 3 and 4. (Classifications of Causes of Death. See also *Manual of Causes of Death,* based upon revision by the International Commission, Paris, 1920, published for the Bureau of the Census by the Government Printing Office, 1924.)

Bowley, A. L., *An Elementary Manual of Statistics*, part I, chap. 7. (Discussion of Sampling, which requires careful preliminary analysis and classification.)

Chapin, F. S., *Field Work and Social Research*, chap. 5. (Sampling.)

Consult references in Chapters XIV and XV on the collection and tabulation of statistical data.

[Details as to publisher and date of publication of the readings and references given at the close of each chapter are found in the alphabetical list in Appendix A.]

CHAPTER V

THE FREQUENCY DISTRIBUTION

BEFORE presenting the technical methods of grouping quantitative data, some fundamental concepts should be emphasized. *Absolute measurements can never be made and countings are frequently only approximations.* Statistical methods are concerned with estimates as well as with measurements and enumerations. Therefore, the attitude of the statistical expert differs from that of the accountant who works out a perfect balance sheet. There is no way of knowing exactly how many bushels of wheat or tons of coal are produced annually in the United States. We do not know the population of New York State for July 1, 1924, because over four years have elapsed since a count was made.

Accuracy is a relative term in statistics. Even in the physical sciences, where much greater precision can be attained than in the social sciences and where standards of accuracy are much more refined, it is recognized that absolutely exact measurement is impossible. The most delicate instruments merely make possible a greater degree of precision.

A standard of accuracy is essential. This standard should be worked out in advance for each item, whether measured, counted or estimated, and the recorded values should conform to it. The object is to secure data sufficiently exact for the purposes of the particular investigation, and the standard decided upon may not make necessary the highest degree of precision possible. In presenting the results care should be taken to explain how closely the facts conform to the established standard of accuracy. The exactness of the final results of an investigation is not determined by that of the most accurate item. For example, weekly earnings of workers may be secured from payrolls accurate to the cent, but it would be useless to state their annual incomes with this degree of precision, because usually we do not know how much time was lost during the year. *Greater precision of statement does not necessarily mean greater accuracy.*

Besides, a high degree of precision, though attainable, may be a waste of effort. The average annual income of workers accurate to cents and mills would have no practical value in describing their standard of living or in comparing the standards of different groups. It would be folly to weigh a ton of coal with the same precision as the jeweler weighs precious stones. *It is relative, not absolute, accuracy which is important.*

A variable is a measurable attribute the values of which differ from each other, as the ages of the population. These values are distributed along a scale from lowest to highest, or *vice versa,* with more or less regularity. The individual magnitudes differ from each other by unequal amounts. *Variation is fundamental in statistical analysis.*

CLASSIFICATION OF QUANTITATIVE DATA

Masses of data are without meaning and interest until arranged in some logical order or sequence. It is the purpose of this chapter to describe methods of bringing order out of the confusion of many individual values. For example, at the time of entering Columbia College, the Freshmen are given a physical examination, and their height, weight, and other measurements are carefully recorded on individual cards. *The records of one thousand students of the same age were selected at random from the files. The weights, stated to the nearest tenth of a pound, were transcribed.*

These recorded weights without orderly arrangement are just so many figures, dry as dust. What is the typical or most common weight? How many are abnormally under- or over-weight? Would the records of a second thousand of the same age reveal a different average or representative weight? How would the typical weight of girls of the same age compare with that of boys? How does the average weight change from year to year as the individuals increase in age? At a given age can we speak of a usual or normal weight for each height which may be a useful measure with which to compare the weight of a particular individual? These are important questions.

The array. The isolated and ungrouped records of weight yield answers to none of these queries. A first step in bringing order out of confusion is to rank the values in order of size. *This arrangement of ungrouped data according to magnitude is called an array,* and is illustrated in an abridged[1] form by Table 1.

TABLE 1. AN ARRAY OF ONE HUNDRED WEIGHTS

91.8	115.0	120.9	125.8	129.0	132.6	137.1	142.1	149.1	159.8
98.9	115.8	121.3	126.1	129.8	133.0	137.8	142.6	149.9	161.0
103.0	116.3	122.0	126.2	129.9	133.4	138.1	143.1	150.7	163.2
106.7	116.7	122.5	126.8	130.0	134.0	138.7	144.0	151.3	165.1
109.2	118.0	122.5	127.0	130.2	134.2	139.1	144.3	152.2	167.4
110.8	118.8	123.3	127.0	130.8	134.3	139.8	145.1	153.3	170.0
111.8	119.2	124.0	127.7	131.0	135.1	140.0	146.0	154.0	173.4
112.6	119.9	124.5	128.0	131.4	135.3	140.7	146.3	155.1	178.8
113.9	120.0	124.6	128.2	131.9	136.1	141.2	147.4	156.3	185.3
114.3	120.3	124.8	128.7	132.0	136.8	141.8	148.0	157.9	189.9

[1] The one hundred items are selected from the larger sample already described.

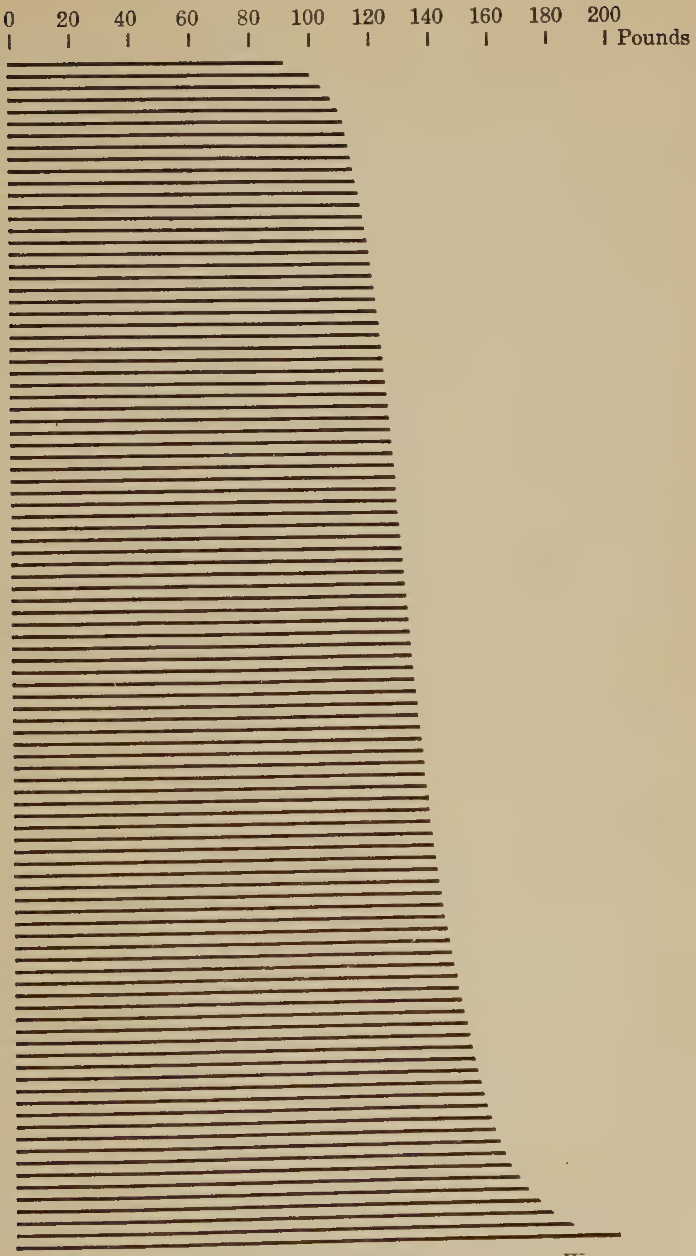


FIG. 3. GRAPHIC REPRESENTATION OF AN ARRAY OF WEIGHTS
(Data from Table 1.)

Inspection of the *array* reveals the range of variation from the lowest to the highest value, about one hundred pounds. Over this range the hundred items are distributed, with a distinct tendency to mass around a central or typical weight. The characteristics of the mass of data begin to appear. Small differences from the central value are more frequent than large ones.

Figure 3 represents the hundred weight items of Table 1 by horizontal lines each drawn the proper length on the scale. Toward the center of the array the lines are of nearly equal length. The differences in the lengths of adjacent lines represent differences in the weights of individuals. These inequalities are greater at the extremes than near the center of the array.

Magnitude classes. The structure of the distribution of the weight items will be more evident if individual values are combined into groups. The next step after the formation of the array is to divide the entire range into *intervals of equal size, called class-intervals*, each with a definite upper and a lower limiting value. How many of these intervals should we have? It will aid in our decision if we experiment with the actual data. First we divide the range from 90 pounds to 210 pounds into *twenty-four intervals of five pounds each*, as 90 and less than 95, 95 and less than 100, 100 and less than 105, and so on. Then all the weights are grouped together which fall between the limits of each successive class until the 1000 items are distributed. All the items in a given class may be regarded as having the mid-value of that class, as 92.5 pounds, 97.5 pounds, 102.5 pounds, which is equivalent to assuming an even distribution over the five-pound interval and the mid-value as the average of all the values within that class.

The actual procedure of grouping the individual items into *class-frequencies*, the number of cases in each class, is of practical importance. If each value is recorded on a separate card, it is simple to sort these cards into five-pound groups and to count the numbers in each group. If the items are to be classified from a list the following method is useful.

TABLE 2. WEIGHTS GROUPED IN FIVE-POUND INTERVALS

CLASS LIMITS (pounds) (1)	MID-VALUE (pounds) (2)	INDIVIDUAL ITEMS (tabulated from list) (3)	FREQUENCY IN EACH CLASS (4)
90 and under 95	92.5	𝍸 /	6
95 " " 100	97.5	𝍸 //	7
100 " " 105	102.5	𝍸 ////	10

Each item is checked in column (3) beside its proper class by a line as indicated, each four vertical lines being crossed by a diagonal line for the fifth item. These are easily summarized in column (4), which gives the number of items grouped in each class-interval.

The frequency distribution. *The array* has now been transformed into a *frequency distribution.* All the cases within any class limits constitute the frequency of that class, whereas the array showed the original individual values. The frequency table is a simplification of the original data by grouping values which are sufficiently alike. It may be defined as *an arrangement of quantitative data in order of magnitude, grouped by a selected class-interval of value so as to reveal clearly the internal structure of the mass of facts for the purpose in view, and so as to be accurate and useful for purposes of summarization, comparison, and analysis.*

Intervals of different sizes may be employed in grouping the data. As a further experiment we shall divide the entire range into twelve ten-pound intervals, instead of twenty-four five-pound classes, as 90 and less than 100, 100 and less than 110, 110 and less than 120. Now the value of each item grouped in a given ten-pound interval is assumed to be at the mid-value of that interval. The frequencies are easily obtained by combining two five-pound classes of the previous tabulation. The range also has been divided into eight fifteen-pound intervals. The frequency distributions for the three different class-intervals are shown in Table 3.

TABLE 3. WEIGHTS OF ONE THOUSAND FRESHMEN

A			B			C		
CLASS LIMITS (pounds) (1)	MID-VALUE (2)	NO. (3)	CLASS LIMITS (pounds) (1)	MID-VALUE (2)	NO. (3)	CLASS LIMITS (pounds) (1)	MID-VALUE (2)	NO. (3)
90 and under 95	92.5	6	90 and under 100	95	13	90 and under 105	97.5	23
95 " " 100	97.5	7	100 " " 110	105	28	105 " " 120	112.5	164
100 " " 105	102.5	10	110 " " 120	115	146	120 " " 135	127.5	370
105 " " 110	107.5	18	120 " " 130	125	245	135 " " 150	142.5	277
110 " " 115	112.5	65	130 " " 140	135	242	150 " " 165	157.5	114
115 " " 120	117.5	81	140 " " 150	145	160	165 " " 180	172.5	39
120 " " 125	122.5	111	150 " " 160	155	89	180 " " 195	187.5	11
125 " " 130	127.5	134	160 " " 170	165	46	195 " " 210	202.5	2
130 " " 135	132.5	125	170 " " 180	175	18			
135 " " 140	137.5	117	180 " " 190	185	9			
140 " " 145	142.5	85	190 " " 200	195	3			1000
145 " " 150	147.5	75	200 " " 210	205	1			
150 " " 155	152.5	54						
155 " " 160	157.5	35						
160 " " 165	162.5	25			1000			
165 " " 170	167.5	21						
170 " " 175	172.5	13						
175 " " 180	177.5	5						
180 " " 185	182.5	5						
185 " " 190	187.5	4						
190 " " 195	192.5	2						
195 " " 200	197.5	1						
200 " " 205	202.5	0						
205 " " 210	207.5	1						
		1000						

Advantages of grouping. The various groupings in Table 3, A, B, and C, show the arrangement of weights within the entire range of 120 pounds in simpler form than the array. Each indicates a concentration at or near some typical weight not yet exactly determined. Regularities are revealed in the grouping about the central value which the ungrouped data do not show. Each tabulation emphasizes the fact that wide deviations from the central value occur less frequently than small ones. Furthermore, comparison with other similar series of facts is made possible, as weights of students in different institutions located in other sections of the country.

How shall we decide upon the class-interval which is to be used in grouping the data? The object of grouping similar values together as if they were alike is to secure the advantages suggested without sacrifice of the necessary accuracy of analysis and summation. Errors sometimes arise from the assumptions in this process of simplification.

Assumptions underlying the formation of frequency-classes. If we wish to calculate an arithmetic average from the ungrouped weights, the separate items need only be added and divided by 1000. The resulting quantity may be termed the *true average weight* (134.41 pounds). If a similar calculation is made from the data of Table A, all individuals in a given class are regarded as weighing the same amount, average value or mid-value of that class. This is done on the assumption of an even distribution of the items between the limits of the class or a concentration of all the items at the mid-value, which assumes an equal number of items located above and below this mid-value.

The total weight of the cases in a given class is secured by multiplying the mid-value by the frequency or number of cases, for example, in Table A, 92.5 pounds times 6, 97.5 pounds times 7, etc. If the weights were ungrouped the actual separate weights of the six students falling between 90 and 95 pounds would be added, whereas in the above grouping all six are regarded as having the same weight (92.5 pounds), and it is only necessary to multiply this value by 6. The two results are not exactly the same, but *approximately* so.

If we compute the arithmetic average from grouped data in five-pound intervals, by methods which will be elaborated in Chapter VI, the result is 134.45 pounds, compared with 134.41 pounds when computed from the separate items ungrouped. The difference is insignificant. Therefore, grouping in five-pound intervals does not destroy the accuracy of the resulting average.

Careful inspection of the two frequency distributions in Tables A and B will suggest the *limitations on the assumption of even distribution over* *a class-interval.* In Table B the ten-pound interval, 90 to 100 pounds, which has 13 items, combines 2 five-pound intervals of Table A. The lower five-pound group, 90 to 95 pounds, has 6 items and the upper five-pound group has 7 items. This is almost even distribution above and below the mid-value of the ten-pound interval. How is the frequency in the next ten-pound interval, 100 to 110 pounds, distributed? There are 28 items, and if these were evenly distributed, 14 items would be located above and the same number below the mid-value, 105 pounds. As a matter of fact, the lower five-pound group, 100 to 105 pounds, has 10 items and the upper group, 105 to 110 pounds, has 18 items. This tendency of the items to mass toward the upper limit of each class-interval continues up to 130 pounds. The opposite tendency appears in the interval 130 to 140 pounds, where the group below the mid-value, 130 to 135 pounds, has 125 items, while the upper group, 135 to 140 pounds, has only 117 items. In the class-intervals above the central value of the entire distribution, the tendency is for the items to mass toward the lower limit of the interval. *The common tendency is to mass toward the central value of the entire distribution.*

The size of the class-interval. When the ten-pound grouping is used, the resulting average is 134.45 pounds, the same to the second decimal as for the five-pound grouping. It is obvious that the ten-pound grouping requires less work in computing the average. On this account it is an advantage to have fewer groups, provided the assumption of even distribution does not lead to significant errors in calculations.

It is clear that the assumption of mid-values in a five-pound interval is closer to the facts of the ungrouped items than a similar assumption for a ten-pound interval. Why, then, is there so little difference in the averages computed from the two distributions? It is due mainly to the fact that the tendency of the items below the average to mass toward the upper limit of each class-interval is balanced by the opposite tendency above the average.

As a third experiment in the effect of the size of the class-interval upon the average, Table C may be used with a fifteen-pound interval. The assumption of even distribution over a fifteen-pound interval is still less in accord with the actual distribution of the individual values. However, this assumption has less influence upon the average than one would expect. Calculating the average in the same manner as before, we find it to be 134.49 pounds, as compared with 134.45 pounds for both the five-pound and ten-pound intervals, and 134.41 pounds for the ungrouped data.

The fifteen-pound grouping, besides producing about one tenth of a pound difference in the average as compared with the ungrouped data, does not describe in sufficient detail how the items are distributed above and below the average. The importance of this detail will appear in later chapters. *The object of grouping is to render the data as simple and compact as possible without sacrifice of the desired accuracy.* As between the five-pound and the ten-pound interval, the latter would seem to have the advantage, because it is simpler to handle and is sufficiently accurate. *In deciding on the size of interval for any given type of data, it is desirable to experiment.* The smaller the interval the larger the number of classes and the more nearly does the assumption of mid-values approximate the actual values, but the greater becomes the labor of handling the groupings. On the other hand, the larger the interval the smaller the number of groups, and the easier they are to manipulate. *Less than ten intervals are rarely satisfactory, and more than twenty are usually unnecessary to attain the desired detail and accuracy.* In fact experiment will show that narrowing the interval and increasing the number of groups beyond a certain point yields no advantage and only imposes added drudgery.

Importance of uniform class-intervals. *The class-intervals of a frequency distribution should be uniform in size throughout the range wherever possible.* In case this is not possible or desirable in the particular distribution, the intervals should be so subdivided as to make combination into uniform intervals practicable. Likewise, undistributed extreme values should be avoided, as a distribution of earnings which bunches the numbers earning under $15 and also those over $30 per week and has uniform class-intervals between these values. The fact that the items at the margins are scattered and that space for publication is limited, frequently makes such grouping necessary, but this grouping interferes with the application of certain statistical methods.

Sometimes the particular data and their uses logically require different sized intervals at different parts of the range. For example, it is adequate for many purposes to classify the age of the population in five-year groups up to twenty-five years, and in ten-year periods above this age. For health and educational work, however, it is desirable to divide the population into smaller classes. The Federal Census of 1920 subdivided the ages of the population under 21 years, under 1 year, 1 to 4, 5, 6, 7 to 9, 10 to 13, 14, 15, 16 and 17, 18 and 19, 20. This classification furnishes the necessary details for the use of health and school officials, and at the same time permits combination into uniform five-year intervals, under 5 years, 5 to 9, 10 to 14, 15 to 19, 20 to 24. *The general rule in forming classes is to make them as detailed as necessary for the purposes in view, but arranged so as to permit combination into intervals of uniform size if desired.*

Intervals of different size in the same frequency table give wrong impressions as to the regularity of the distribution, as illustrated in Table 4.

TABLE 4. DISTRIBUTION OF PERSONAL INCOMES $4000 to $50,000, UNITED STATES, 1918[a]

INCOME (1)	NUMBER (thousands) (2)
$4,000 and under 5,000	430
5,000 " " 6,000	235
6,000 " " 7.000	143
7,000 " " 8,000	95
8,000 " " 9,000	67
9,000 " " 10,000	48
10,000 " " 11,000	36
11,000 " " 12,000	28
12,000 " " 13,000	22
13,000 " " 14,000	18
14,000 " " 15,000	15
15,000 " " 20,000	47[b]
20,000 " " 25,000	25
25,000 " " 30,000	15
30,000 " " 40,000	17[b]
40,000 " " 50,000	9

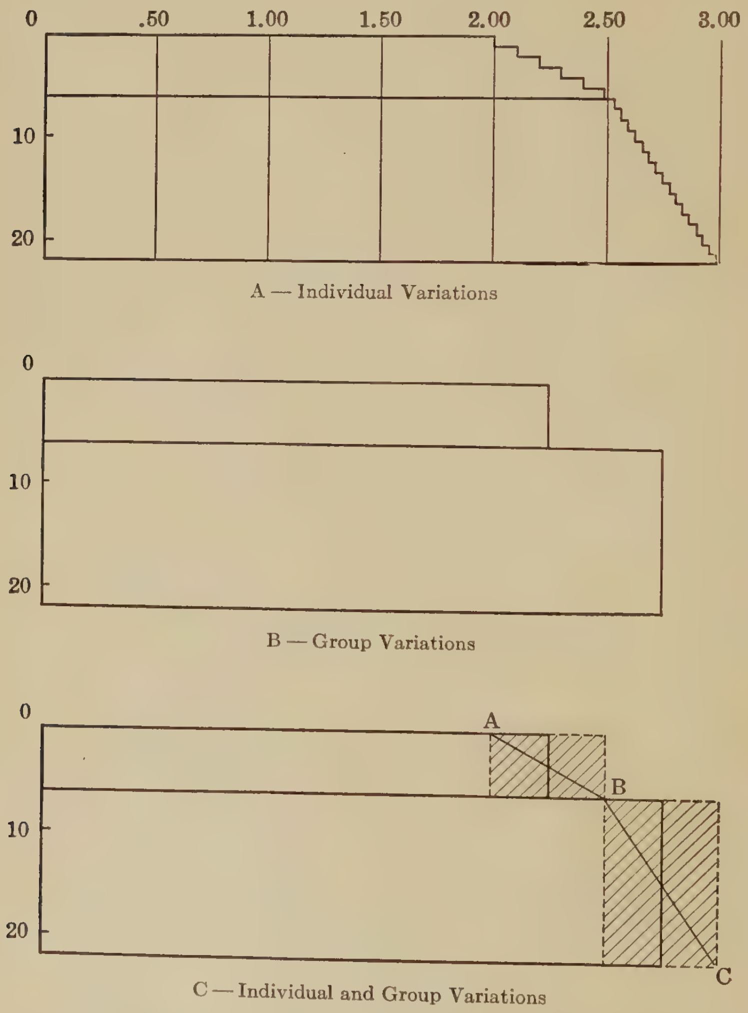
[a] Data compiled from *Income in the United States*, National Bureau of Economic Research, vol. I, p. 133.
[b] Interval of a different size accounts for the increase in the frequency.

THE GRAPHIC REPRESENTATION OF AN ARRAY AND A FREQUENCY DISTRIBUTION

Graphic representation assists the student to understand the assumptions underlying grouped data, and to visualize the procedure involved in transforming an array into a frequency distribution.

The data used for illustration are piece-rate earnings, instead of weights. The individual earnings of 336 workers are grouped in fifty-cent intervals, after having been ranked in order of size in an array from the lowest to the highest values. They are distributed over the range of value with a fair degree of regularity. The unit in which wages are paid is not capable of so minute subdivision as was the case in the measurement of weights. This is especially true of wages paid by the day which are not evenly distributed along the scale of values but tend to concentrate at certain values. But the illustration uses piece-rate earnings where the smallest subdivision of the unit of payment is one cent. Therefore, within any class-interval of fifty cents there are many subdivisions possible, and the data may be treated by the same methods as have been described for weight. The array and the frequency distribution are shown in the diagrams, Figures 4, 5 and 6. The frequency table is presented as a part of Figure 5.

The amount of the wage is indicated on the horizontal scale and the



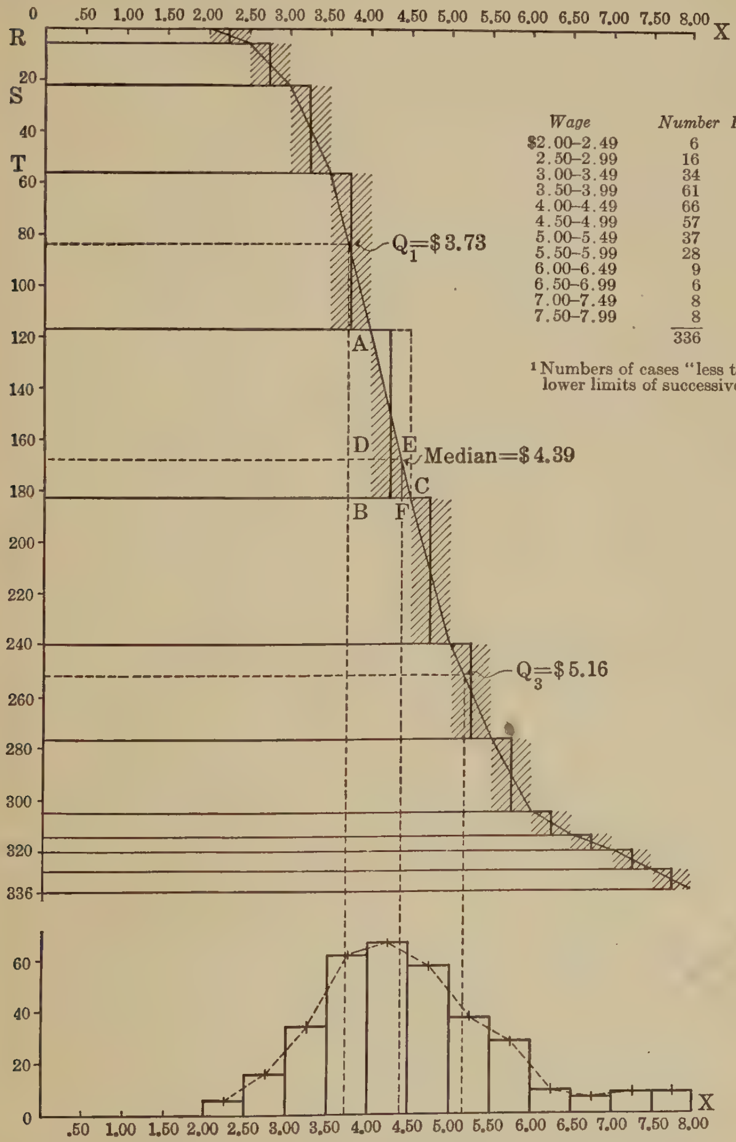
A — Individual Variations

B — Group Variations

C — Individual and Group Variations

FIG. 4. DISTRIBUTION OF PIECE-RATE EARNINGS IN FIFTY-CENT GROUPS

number of wage-earners on the vertical scale, starting with zero on the left and at the top of the diagram. A given vertical distance everywhere represents the same number of workers and a given horizontal distance the same amount of wage.



Wage	*Number*	*Cumulative Frequencies.*[1]
$2.00–2.49	6	0
2.50–2.99	16	6
3.00–3.49	34	22
3.50–3.99	61	56
4.00–4.49	66	117
4.50–4.99	57	183
5.00–5.49	37	240
5.50–5.99	28	277
6.00–6.49	9	305
6.50–6.99	6	314
7.00–7.49	8	320
7.50–7.99	8	328
	336	336

[1] Numbers of cases "less than" the lower limits of successive classes.

FIG. 5. DISTRIBUTION OF PIECE-RATE EARNINGS IN FIFTY-CENT GROUPS
FIG. 6.{ FREQUENCY HISTOGRAM} AND POLYGON

Figure 4 A shows an array of the wage items within each of the first two fifty-cent intervals. Between $2.00 and $2.50 are distributed 6 items, and between $2.50 and $3.00 are 16 items. The individual earnings are arranged *on the assumption of even distribution* by allowing a rectangular step on the horizontal scale and a space on the vertical scale to represent each item. Within a given class-interval these steps progress regularly from the lowest value of that class to the highest, and so on. The distance between the heavy limiting lines on the vertical scale indicates the number of cases in each class-interval, 6 and 16 respectively.

Figure 4 B presents the same facts with the individual variations in value eliminated. The 6 items in the lower class-interval are here given the same value on the horizontal scale, the mid-value of the fifty-cent range, $2.25, *on the assumption of concentration of all items at the mid-value.* Likewise, the 16 items of the second group are concentrated at $2.75, the mid-value of the next higher interval. These mid-values are used in calculating the arithmetic average, exactly as in the case of the weights. Therefore, the rectangular space on the diagram which represents 6 items extends to $2.25 on the horizontal scale, and the space representing 16 items extends to $2.75. The individual step differences are no longer recognized.

Figure 4 C combines the facts of both A and B. The cross-hatched areas show two things: (1) the range along the horizontal scale over which the items of the class-interval are actually distributed from the smallest value to the largest, and (2) the number of items in each group measured on the vertical scale. The diagonal *AB* shows the increasing values in that class of 6 items which was indicated by the steps in Figure 4 A. The diagonal *BC* represents the same for the 16 items of the second class, the steps in each case having been smoothed by the straight lines *AB* and *BC*.

The objection may be raised that values are not usually distributed as evenly as pictured in these diagrams. This was illustrated in the case of weights. The items are likely to occur in larger numbers toward the lower or upper limits of the class — within one half of the interval as compared with the other. This objection merely emphasizes the importance of care in the determination of the class-interval, its width and the position of the upper and lower limiting values. The need for experimentation has already been pointed out. While classification and grouping of the individual values are usually essential for ease and clearness in handling, it is necessary to make the results approximate as closely as possible the true results.

In Figure 5 the procedure of Figure 4 C has been carried out for each group in the entire array of 336 wage items. The table in the upper right hand corner gives not only the frequencies for each fifty-cent interval, but in the last column cumulates these frequencies by adding each successive frequency to the total frequencies in all preceding classes. The vertical scale is made to include 336 items spaced at equal intervals. The distance between the heavy limiting lines on the vertical scale represents the number of items in each class-interval, as does also the vertical cross-hatched area. The cumulative frequencies enable the student in drawing the diagram to locate easily the heavy horizontal lines marking off the limits of the separate classes. For example, a line is drawn through R at 6 on the vertical scale, a second line is drawn through S at 22 $(6 + 16)$, a third line is drawn through T at 56 $(6 + 16 + 34)$, and so on. The total vertical length of the cross-hatched areas represents the total of the items, 336. In each group the width of the cross-hatched area is the same, measuring on the horizontal scale the uniform fifty-cent interval.

In Figure 6 the frequency histogram and polygon are represented. The horizontal scale is repeated at the bottom for Figure 6 exactly as it is at the top of Figure 5. All the cross-hatched areas are allowed to fall upon the common base line $O\,X$ of Figure 6, preserving the same horizontal and vertical dimensions as in Figure 5, and locating them in exactly the same positions on the horizontal scale. The result is to place these rectangular areas side by side at their proper values on the horizontal scale, representing by their respective heights above the base line the frequencies in each class. *This is a surface of frequency, called a histogram or column diagram.* The total area represents 336 wage items, and is identical with the total cross-hatched area of Figure 5. By the dotted line connecting the mid-points of the tops of the columns the histogram is converted or smoothed into the *frequency polygon.* The polygon may be further smoothed by appropriate methods into a *frequency curve. These forms by their contours and areas picture the distribution of values around a central value. This procedure is fundamental in statistical analysis of some types of data.* Much of statistical method has for its object the description of frequency distributions in summary form and the comparing of one distribution with another.

The student should carry out this entire experiment using Table 3 *A or B, the frequency distribution of weights.* It should be remembered that the object of this experiment is to make clearer the transition from an array of values to a frequency grouping, and to emphasize the assumptions underlying this grouping. Figure 5 will be used again in Chapter VII to exemplify the graphic location of the median and quartiles, which accounts for certain details on the diagram not yet explained. The reader is asked to disregard these for the present.

The *histogram or column diagram* and the *frequency polygon* can be plotted [1] direct from the frequency table without the more elaborate procedure just explained. In actual statistical practice this is done. The method will be illustrated in the next section.

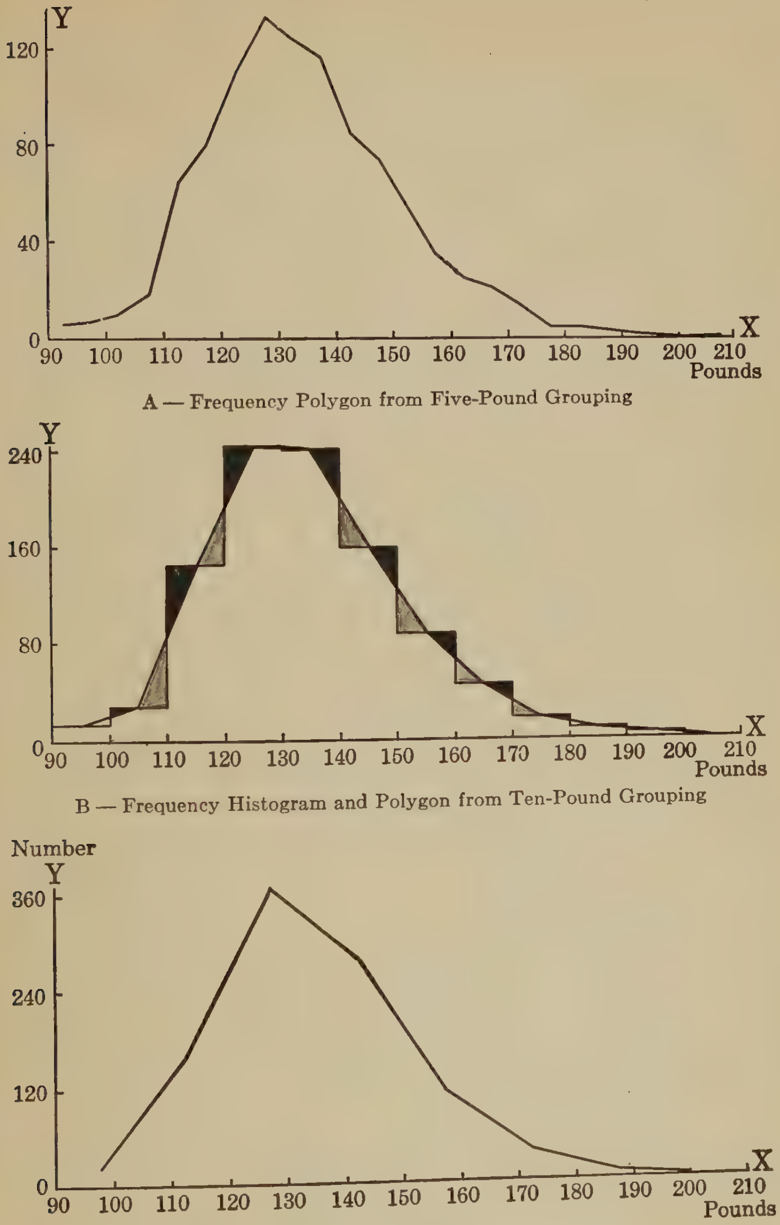
THE EFFECT OF THE SIZE OF THE CLASS-INTERVAL UPON THE POLYGON AND HISTOGRAM

In Table 3, page 57, different groupings of weights in intervals of five, ten and fifteen pounds were presented. These frequency distributions are portrayed in Figures 7 A, B, and C.

The polygons of Figures 7 A and C are plotted by a simpler method than that used in Figure 6. The amounts of weight are indicated by distances on the horizontal scale at the bottom of each diagram. The base lines of each are divided into equal intervals. The frequencies are indicated by distances measured on the vertical scales at the left of the diagrams. Figure A is drawn by plotting the frequencies of Table 3 A the proper vertical distance above the mid-points of successive five-pound class-intervals. These vertical distances, or ordinates, are indicated by dots located at the top of each ordinate. Straight lines are drawn connecting the dots and forming the *frequency polygon.* Likewise, Figure C pictures the frequencies of Table 3 C in fifteen-pound intervals, which are represented by distances on the base line three times as great as for the intervals of Figure A. Likewise, the number of cases indicated by a given distance on the vertical scale of A must be trebled for C, in order to portray the greater frequencies of the several classes resulting from combining into larger intervals. Dots are located directly above the mid-points of successive fifteen-pound intervals and these are connected by straight lines, as in A. *The respective areas of Figures A, B, and C are equal, representing one thousand cases.*

Figure B is drawn so as to show both *histogram* and *polygon,* the latter superimposed upon the former. The class-interval is ten pounds, indicated by a distance on the base line twice as great as that of A. The number of cases indicated by a given distance on the vertical scale of A must be doubled for B. The histogram is plotted from the frequencies

[1] The question may be raised as to the inclusion of zero or the origin on the horizontal scale in plotting a frequency distribution. In Figures 6 and 8 of this chapter it has been included, but in Figures 7 and 9 it has not been shown, because the lowest values on the scale are located far from zero. The importance of including it wherever possible will be discussed in Chapter IX. For more detailed discussion refer to Horace Secrist: *Readings and Problems in Statistical Methods,* pp. 385–94.



A — Frequency Polygon from Five-Pound Grouping

B — Frequency Histogram and Polygon from Ten-Pound Grouping

C — Frequency Polygon from Fifteen-Pound Grouping

FIG. 7. REPRESENTATION OF 1000 WEIGHTS GROUPED IN CLASSES OF DIFFERENT SIZES

(Data from Table 3.)

of Table 3 B. The number of cases in each successive class is represented by a rectangular area whose height above the base line is determined from the scale on the left. Only the tops of these rectangles are shown in the diagram, instead of drawing also the lines marking the limits of the classes. This procedure causes the outline of the histogram to appear in steps which rise to the peak of the diagram and then descend in similar fashion. In Figure 6 the rectangles representing frequencies are complete. Either form of the histogram is employed in practice. The polygon is superimposed upon the histogram by connecting the mid-points of the tops of the rectangles by straight lines, as in Figure 6.

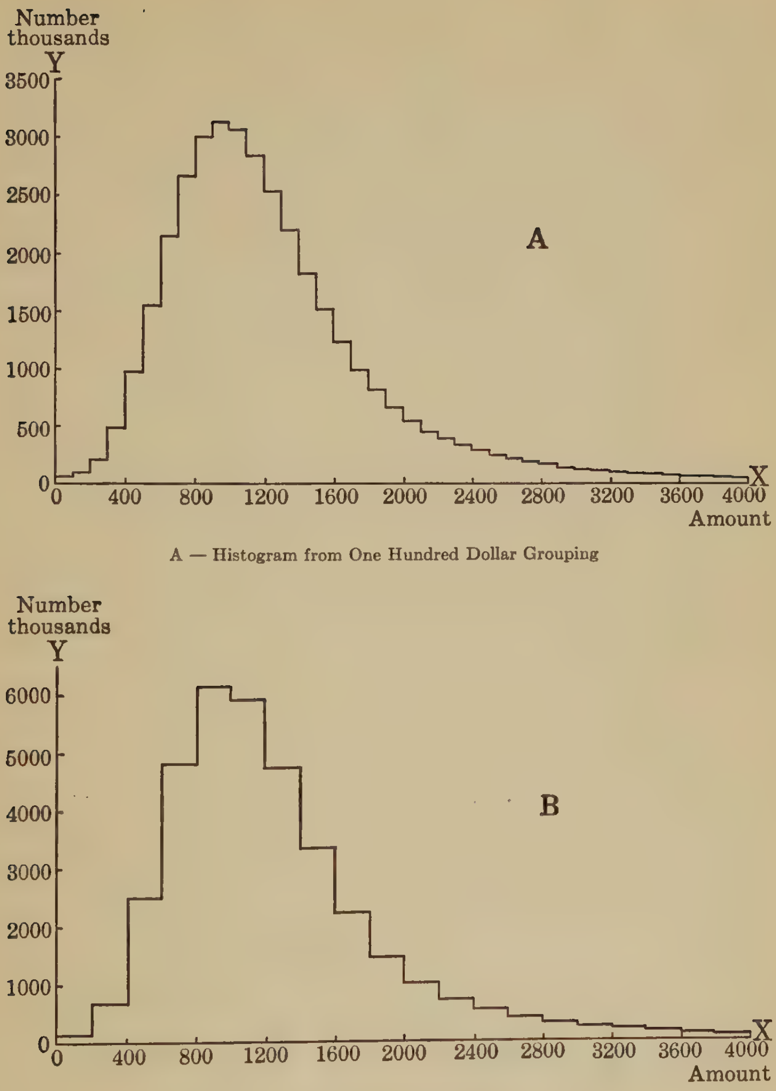
The areas of the histogram and polygon in Figure 7 B are equal. The shaded areas represent the portions of the rectangles cut off by connecting the mid-points, in constructing the polygon. They are equal to the areas included in the polygon which are not a part of the area of the histogram.[1]

The reader will observe that as the width of the class-interval is changed in Figures A, B and C the highest point of the polygon shifts on the horizontal scale. For example, in A the peak is at 127.5 pounds, in B it has shifted to 125 pounds, and in C it moves back to 127.5 pounds. This indefinite position of the highest point of the frequency polygon, due to the particular class-interval employed, will be considered in Chapter VIII in the discussion of the mode as a form of average.

Narrowing the class-interval smooths the histogram or polygon. As the interval is diminished in size, the number of points plotted grows greater, and the graduation between any two becomes less. This procedure presupposes enough cases to permit a greater number of classes and still to maintain a regular frequency distribution. Let us use for illustration data from the field of income statistics in the United States, where the number of cases is very large.

The frequency distribution in $100 classes and the source of the data are given in Chapter VII, page 108. The number of incomes in each class is large, the frequencies varying from a minimum of 63,000 to a maximum of over 3,100,000. The histograms, Figures 8 A and B, are plotted as in Figure 7 B. By narrowing the class-interval from $200 to $100, Figure 8 A, the histogram is smoothed. If the interval were narrowed further and the number of classes were indefinitely increased the form of the resulting diagram would approach a smooth curve, called the *frequency curve*. The general shape and range of this frequency curve representing incomes should be noted carefully. The value at or near which the largest number of items is located and about which the others

[1] G. U. Yule: *An Introduction to the Theory of Statistics*, 6th ed., 1922, pp. 84–87.



A — Histogram from One Hundred Dollar Grouping

B — Histogram from Two Hundred Dollar Grouping

FIG. 8. DISTRIBUTION OF PERSONAL INCOMES UNDER $4000, IN THE UNITED STATES, 1918
(Data from Table 19, Chapter VII.)

cluster is not at the center of the base line, the mid-point of the entire range from 0 to $4000. It is situated at a point about one fourth of the distance from the zero end of the horizontal scale. This is the usual or *normal* form of curve for both incomes and weights.[1]

It will be observed that the highest part of the histogram in Figure 8 A is not identical in location with that of Figure 8 B. The class of greatest frequency shifts slightly when the size of the interval is changed, as in Figure 7. The areas of A and B are the same.

The characteristic distribution of heights. Table 5 presents the heights of the same one thousand Freshmen whose weights have been examined. The grouping is in intervals of one inch, the original measurements having been given *to the nearest tenth* of an inch.

TABLE 5. HEIGHTS OF ONE THOUSAND FRESHMEN

CLASS LIMITS	FREQUENCY	CLASS LIMITS	FREQUENCY
(inches)	(f)	(inches)	(f)
(1)	(2)	(1)	(2)
60 and under 61	3	68 and under 69	154
61 " " 62	12	69 " " 70	115
62 " " 63	21	70 " " 71	67
63 " " 64	55	71 " " 72	45
64 " " 65	51	72 " " 73	29
65 " " 66	135	73 " " 74	13
66 " " 67	139	74 " " 75	8
67 " " 68	152	75 " " 76	1
		Total	1000

The frequency polygon is portrayed in Figure 9 A.

In the data as presented certain irregularities in the frequencies appear. For example, in the class 64 to 65 inches there are 51 cases which are less than the number in the preceding class. This causes a zigzag in the polygon, Figure 9 A. The two middle classes have almost exactly the same frequencies, 152 and 154 respectively. The arithmetic average height, computed from Table 5, is 67.6 inches, which is almost in the middle of the entire range as measured on the horizontal scale, and is located at about the mid-value of the class having 152 cases. *The irregularities in the frequencies are due to the fact that not enough cases have been measured.* With these irregularities smoothed out, as they would be for a very large number of heights, the resulting frequency curve is of the type called the *bell-shaped symmetrical*, Figure 9 B. This is the usual or *normal* type for height and is in contrast to those for weight and income.

[1] F. L. Hoffman: *Army Anthropometry and Medical Rejection Statistics*, pp. 32–42, Prudential Insurance Company, 1918. Compare curves of weight and height of army recruits.

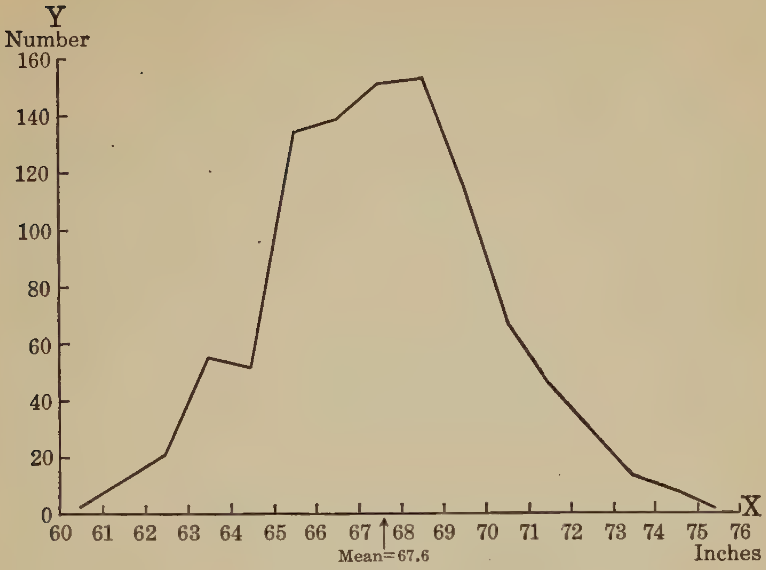


FIG. 9A. DISTRIBUTION OF 1000 HEIGHTS IN ONE-INCH CLASSES
(Data from Table 5.)

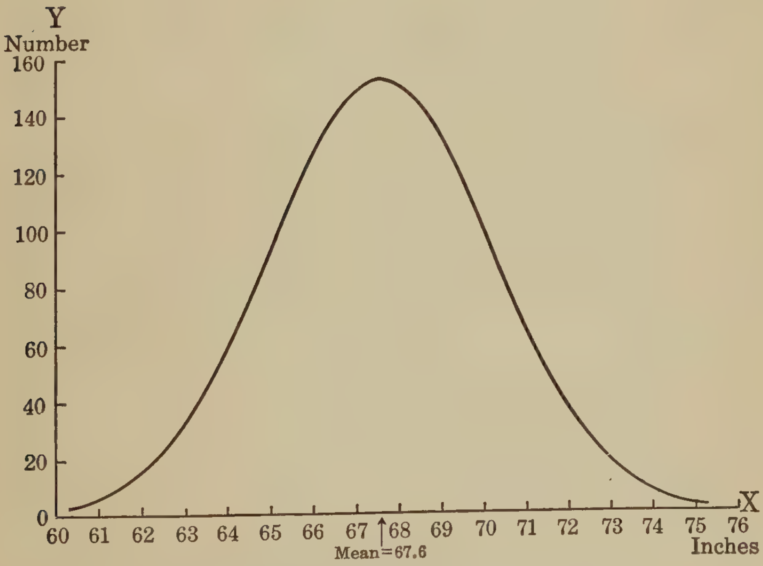


FIG. 9B. THE IDEAL FREQUENCY CURVE FOR HEIGHTS —
THE BELL-SHAPED SYMMETRICAL TYPE

THE BELL-SHAPED SYMMETRICAL CURVE

The understanding of this type of curve is important for the chapters which follow. In them we shall discuss measures of central tendency and measures of variability used in describing frequency distributions. The bell-shaped symmetrical curve is characterized by the symmetrical arrangement of the items around the central value. As in the case of all frequency distributions, the small deviations from the central or typical value occur most frequently and the larger deviations less frequently the greater the distance from the central value. In the bell-shaped curve the total number and amount of deviations above the average or type value are equal to the number and amount of deviations below. The average is located in the middle of the range — midway between the highest and lowest magnitudes. If the part of the area under the curve above the average is superimposed upon the part below the average, the two areas coincide.

Different distributions depart in varying degrees from perfect symmetry,[1] depending upon the more or less frequent occurrence of extreme items either at the higher or at the lower values on the scale. Both weight and income are examples of moderately asymmetrical distributions, the extreme variations occurring toward the high values on the scale. In some arrangements of data the asymmetry extends toward the lower values. In the latter case the typical value will be located above the center of the range, whereas, in the former, the average is located below the mid-point of the entire range from lowest to highest magnitude.

CONTINUITY OF VALUES ALONG THE SCALE

The array and grouping of data raise certain fundamental queries. How regular and continuous is the distribution of items over tho successive values on the scale? Do *gaps* occur where there are no values recorded? If so, is this because we decide to measure weight only to the nearest quarter-pound, or age to the nearest birthday? Or, have we included too few cases to fill the gaps or to smooth the irregularities in the frequencies? Is there a type of data in which values occur at certain points on the scale and not elsewhere? Are these values sometimes determined by custom in fixing the units of measurement? *Cautions are necessary in handling different kinds of data in which these gaps occur.*

Discrete Data. A distinction must be drawn between *discrete* and *continuous* data in frequency distributions. In discrete data *gaps* occur

[1] G. U. Yule: *An Introduction to the Theory of Statistics*, chap. VI, and E. L. Thorndike: *Mental and Social Measurements*, chap. III, contain good examples of frequency distributions of different types.

in the measurements regardless of the number of cases. The intervals in which the values are recorded are not divisible, as they are in weight, where the measurements may be made to the nearest quarter or eighth of a pound, or to an even finer unit of value. *In discrete data the record is made with complete accuracy at a definite position on the scale.* For example, suppose we wish to group the classes in a large university according to the number of students in each, giving the number of classes of each specified size. It is not possible to have a fractional interval; a class has 15 students not 15½. There are no cases except at integral amounts. Again, if an employer pays time wages of $5 per day and there are no additions for overtime or deductions for spoiled work, the earnings at the end of a week of five and one half days must be $27.50. No number, however large, of similar wage records could produce a distribution so continuous along the scale as the measurements of weight. Business practice often limits the number of subdivisions on the scale. There are gaps where cases do not occur. Data of this character may be termed *discrete.*

In Figure 5 piece-rate earnings are used. In this case the smallest unit of payment is one cent. Therefore, the data, classified in fifty-cent groups, are distributed with a fair degree of regularity over the scale. In any class-interval fifty subdivisions are possible. In handling this distribution the same assumptions may be made and the same methods may be employed as for weights and heights.

Continuous data. *Statistical records do not necessarily measure all possible magnitudes.* For example, natural phenomena, as height, weight and age, are not recorded with complete accuracy. Any specific value does not mean a single point on the scale, but *a range between two limits.* The original class-interval, according to which the measurements are made, is determined frequently by the requirements of the investigation, as height *to the nearest quarter-inch*, weight *to the nearest quarter-pound*, age *to the nearest birthday in years.* In all these cases the units of magnitude may be further subdivided and values could be stated. Individual differences do actually exist but may not be recorded. Items are considered alike which if more exactly measured would appear different. For example, measurement to the nearest quarter-inch allows an eighth of an inch variation above or below the value actually stated. All cases falling within this range of an eighth above or below are recorded at the specific quarter-inch. Values are usually further grouped for study and analysis. The weight data used in this chapter as an example were originally recorded to the nearest tenth of a pound. Afterward they were grouped into five- or ten-pound intervals. Provided enough cases are included gaps do not occur in this type of data and the series is designated *continuous.*

In this kind of material the items when grouped may not be evenly distributed over all possible values within the class, but at least the items are not massed at specific values within the class. Therefore, in determining the class-interval for such distributions, *the width of the interval* is of greater importance than the exact points at which the upper and lower limits and the mid-value of the class are located. If the interval is made too wide in range the assumption of even distribution departs too far from the facts.

Suggested experiment. With an ordinary ruler marked in eighths of an inch, a large number of some natural objects, as leaves or bean pods, may be measured to the nearest quarter-inch of length. The objects should be divided among several individuals in order to secure a larger total number of measurements in a reasonable time. Greater speed can be attained if each person arranges his objects as nearly as possible in ascending order of length before beginning the measurements. The records should be made only in quarter-inches, which allows a variation of about an eighth above or below the value actually recorded. Finally, each person should make a frequency table of the number of items he has measured at each quarter-inch. These frequencies can easily be combined for each quarter-inch interval into a general frequency distribution for all the objects measured. Each person should compute an arithmetic average length for his own objects and compare it with the average for the entire number.

Now, let each person exchange his objects with another, in order that the same objects may be measured a second time but by a different person, in exactly the same manner as before. The individual and combined frequency distributions are different. But the two averages obtained from the general frequency distributions representing the measurements of all the objects by two different persons are surprisingly alike. Table 6 represents the results of such an experiment in measuring the length of 469 bean pods to the nearest quarter-inch, each object having been measured by two different persons.

The measurements recorded at 12 quarters actually ranged from $11\frac{1}{2}$ quarters to $12\frac{1}{2}$ quarters; those recorded at 13 quarters ranged from $12\frac{1}{2}$ quarters to $13\frac{1}{2}$ quarters, etc. The values entered in column (1) are the mid-values of a quarter-inch class-interval. Although the combined frequencies secured from first and second measurements of the same objects show decided differences, the two averages differ by only one tenth of a quarter-inch.

[illegible]

TABLE 6. MEASUREMENT OF BEAN PODS BY DIFFERENT PERSONS — CONTINUOUS DATA

CLASS-INTERVAL QUARTER-INCH MID-VALUES (1)	COMBINED FREQUENCIES FIRST MEASUREMENT (2)	COMBINED FREQUENCIES SECOND MEASUREMENT (3)	DIFFERENCES IN FREQUENCIES (4)
12 quarter-inches	2	1	1
13 " "	7	7	0
14 " "	25	30	5
15 " "	40	46	6
16 " "	63	61	2
17 " "	92	91	1
18 " "	85	83	2
19 " "	81	80	1
20 " "	56	52	4
21 " "	13	12	1
22 " "	4	6	2
23 " "	1	0	1
	469	469	

Average length from all first measurements 17.5 quarter-inches.
Average length from all second measurements 17.4 quarter-inches.

The value of such an experiment may be summarized:

(1) It affords a first-hand example of *continuous data*. The records could have been made in eighths of an inch.

(2) It shows that an average from a small number of measurements gives a very indefinite notion of typical length. The average obtained from each individual's measurements differs more or less widely from the average for the entire number of objects. It emphasizes the importance of measuring a considerable number of cases.

(3) It indicates the difficulty of accurate observation. Two individuals measuring the same objects secure different results.

(4) It emphasizes the fact that *errors of observation may be cancelled by averaging the measurements*. Some second measurements are lower and some higher than the corresponding first measurements. These differences have an insignificant effect upon the average, since the resulting errors tend to balance. This kind of error is described as *accidental, unbiased, or compensating*.

But errors are not all of this type. The markings on one of the rulers may be wrong. If so, all measurements made with this ruler are in error and all the errors are in the same direction. Averaging a large number of these measurements would not eliminate the errors. They are described as *biased, constant, or cumulative.* A fuller discussion of the kinds of error will be found in Chapter XI.

POSITION OF THE LIMITS OF THE CLASS

The one thousand weights may be grouped in a manner different from that in Table 3 A, retaining the five-pound interval. The mid-values may be stated in integers — 90, 95, 100; and the class limits may be given in fractions — 87.5 to 92.5, 92.5 to 97.5, 97.5 to 102.5. This method of grouping makes calculations easier because fractions in the mid-values are avoided; but it proves less convenient and accurate in tabulating the individual items because of the fractional limits to each class. The arithmetic average calculated from the grouping just suggested is 134.35 pounds, compared with 134.45 pounds, computed from either the original five-pound or the ten-pound groupings. The differences are insignificant. For data of this continuous type the specific position of the class limits and of the mid-values is not so important as the width of the interval of classification. *It is always important to state the class limits in such a manner as to make the classes mutually exclusive.*

On the other hand, let us examine data in which there exists a distinct tendency for the items to concentrate at certain values. Five hundred grades, received at the entrance examinations in English I (grammar, composition, and reading), by applicants for admission to Columbia College, are classified in Table 7 by single per cents.

The concentration at multiples of five, indicated by the numbers in bold-faced type, is especially evident in English grades. If five hundred marks in plane geometry are classified in the same manner, the distribution will be much more regular. In an exact subject such as mathematics finer distinctions and more accurate appraisals can be made and, therefore, the tendency of the grades to concentrate at certain values along the scale becomes less.

If it is decided to use a five-per-cent interval in grouping these five hundred English marks, the determination of the positions of the upper and lower limits of the classes and the positions of the mid-values is of first importance in the interest of accuracy. With the facts in mind concerning concentration, the logical procedure, if possible, is to locate the mid-values of the classes at these points of concentration, 20, 25, 30, 35 per cent, as in Table 8. This method of grouping assumes that concentration is as likely to be upward as downward. In some types of data, as age statistics, the concentration is downward and it would be better to locate the massing points at the lower limits of the groups.

TABLE 7. FIVE HUNDRED MARKS IN ENGLISH CLASSIFIED BY SINGLE PER CENTS

GRADE PER CENT (1)	FREQUENCY (2)	GRADE PER CENT (1)	FREQUENCY (2)
20	20	52	10
21	0	53	3
22	1	54	3
23	1	55	20
24	0	56	0
25	20	57	1
26	0	58	4
27	0	59	0
28	0	60	25
29	1	61	3
30	38	62	13
31	0	63	8
32	3	64	2
33	3	65	15
34	3	66	0
35	47	67	2
36	1	68	6
37	0	69	0
38	9	70	19
39	2	71	1
40	53	72	2
41	0	73	0
42	4	74	0
43	2	75	10
44	2	76	0
45	55	77	1
46	0	78	1
47	5	79	0
48	18	80	7
49	0	85	3
50	46	90	3
51	4	Total	500

TABLE 8. CLASSIFICATION OF FIVE HUNDRED MARKS IN ENGLISH BY FIVE-PER-CENT INTERVALS

CLASS LIMITS (per cent) (1)	MID-VALUE (per cent) (2)	FREQUENCY (3)	CASES CONCENTRATED AT MID-VALUES (4)
17.5 and under 22.5	20	21	20
22.5 " " 27.5	25	21	20
27.5 " " 32.5	30	42	38
32.5 " " 37.5	35	54	47
37.5 " " 42.5	40	68	53
42.5 " " 47.5	45	64	55
47.5 " " 52.5	50	78	46
52.5 " " 57.5	55	27	20
57.5 " " 62.5	60	45	25
62.5 " " 67.5	65	27	15
67.5 " " 72.5	70	28	19
72.5 " " 77.5	75	11	10
77.5 " " 82.5	80	8	7
82.5 " " 87.5	85	3	3
87.5 " " 92.5	90	3	3
		500	381 = 76 per cent

Arithmetic average grade = 47.0 per cent
Average from ungrouped data = 46.9 per cent

Column (4) gives the number of marks in each class actually located at the respective mid-values (see Table 7). The total of such marks is 381, or over three fourths of all the grades. Since the mid-values are used in computing the arithmetic average, and since all values within a class are assumed to have the mid-value, this arrangement of classes produces a high degree of accuracy in obtaining the average. This is demonstrated by the fact that there is no essential difference between the arithmetic average obtained from this grouping and the average computed from the individual ungrouped grades (47.0 per cent as compared with 46.9 per cent).

The effect of inaccurate grouping. The same five-per-cent interval may be used, but with different upper and lower limits. Since marks are stated to the nearest whole per cent, all those stated at 20 per cent really range from 19.5 to 20.5 per cent, and those stated at 24 per cent range from 23.5 to 24.5 per cent. Therefore, the class-intervals may be arranged as follows:

Class Limits (per cent) (1)	Mid-Value (per cent) (2)	Frequency (3)	Cases at Lower Limit of Class (4)
19.5 and under 24.5	22	22	20
24.5 " " 29.5	27	21	20
29.5 " " 34.5	32	47	38
34.5 " " 39.5	37	59	47
etc.		etc.	

It will be observed from column (4) that 20 of the 22 grades are located at the extreme lower limit of the first class; that 20 of the 21 grades in the second class are similarly located. All the grades in the three highest classes are located at the respective lower limits. In the entire frequency table so constructed, over three fourths of the grades would be located at the extreme lower limits of the successive classes. It is evident that an average computed from this distribution, using the mid-values at 22, 27, 32, 37, and assuming all grades in the respective classes to be located at the mid-values, would do violence to the facts. The average would be too high by about 1.5 per cent because the grades which are actually located, to a large extent, at the lower limits would be treated as if they were located at the mid-values. In this manner three fourths of the grades would be raised about half an interval. This grouping must be rejected in the interest of accuracy.

It is not unusual for quantitative data to concentrate to greater or less degree at certain values. In these cases care must be exercised in the grouping, both in reference to the size of the intervals and their limits.

For some purposes it is more important to know the total number of items falling above or below a specific value on the scale than to know the frequency in any single class; for example, the number and percentage of incomes in the United States below $2000. This information is furnished by a *cumulative frequency table* or graphic representation. Figure 5, p. 63, represents this type of arrangement. A fuller discussion of the uses of the cumulative arrangement of data will be presented in Chapter VII.

The array and frequency distribution have been described as steps in reducing original data to a more compact form for purposes of treatment. The forms of the frequency distribution have been shown graphically. In this manner the arrangement about a central value can be made clear. How may a frequency distribution be described by a few values or measures, significant for purposes of comparison and interpretation? The following chapters attempt an answer.

READINGS

Elderton, W. P., and Ethel M., *Primer of Statistics,* chaps. 1, 2 and 3. (An introduction to laboratory practice.)

Mills, F. C., *Statistical Methods Applied to Economics and Business,* chap. 3.

King, W. I., *Elements of Statistical Method,* chap. 11.

Rugg, H. O., *Statistical Methods Applied to Education,* chap. 4.

Pearl, Raymond, *Medical Biometry and Statistics,* chap. 4.

Secrist, Horace, *An Introduction to Statistical Methods,* chap. 5, pp. 144–57.

Zizek, Franz, *Statistical Averages.* Translated by Warren M. Persons, part I, chap. 5.

Jones, D. C., *A First Course in Statistics,* chap. 2.

REFERENCES

Yule, G. U., *An Introduction to the Theory of Statistics,* 6th ed., chap. 6. (Types of frequency distributions presented graphically.)

Thorndike, E. L., *An Introduction to the Theory of Mental and Social Measurements,* 2d ed., chap. 3. (Illustrations of types of frequency distributions.)

—— —— *Individuality,* chaps. 1 and 2.

Introduction to Frequency Curves and Averages, American Telephone and Telegraph Company, Statistical Methods Series No. 1, issued by the Chief Statistician, 1921.

Rietz, H. L. (Editor), *Handbook of Mathematical Statistics,* chap. 2. (This chapter is written by Professor Rietz.)

[Details as to publisher and date of publication of the readings and references given at the close of each chapter are found in the alphabetical list in Appendix A.]

CHAPTER VI

METHODS OF SUMMARIZATION AND DESCRIPTION — THE ARITHMETIC AVERAGE

Unorganized masses of data tell us little or nothing. The same facts arranged in an orderly manner, classified and then summarized, may reveal what we wish to know. The frequency distributions in the preceding chapter are very useful as compared with isolated and ungrouped measurements. They show a central tendency amid diversity of measurements but do not reduce it to a single quantity. From the ungrouped data or the frequency distribution we cannot describe the weight of entering Freshmen by a single value. It is important to know how much more wages men earn than women engaged in the same kind of work, but the frequency tables of their respective wages do not present a typical wage value for men and for women.

The frequency table presents all the items in orderly arrangement, but comparison of one distribution with another or with many others, placed in columns side by side, involves holding before the mind at the same time the details of frequency for each series. There is no single quantity which represents all the measurements of one series — for example wage, weight, or age — for purposes of comparison and description. For a single frequency distribution or a small number of such series, graphs may be prepared showing frequency curves which present the detailed facts of the entire array. Two or more of these curves may be placed side by side or superimposed on the same sheet for purposes of comparison. The difference in the curves may be described in words but not in quantities. Furthermore, the number of such curves which can be placed on the same sheet for comparison without confusion is very limited. For instance, the wages of a score of different establishments could not be compared in this manner.

It is clear that the frequency distribution requires further reduction to one or a few significant quantities, provided always that this is possible without too much sacrifice of accuracy of interpretation. We need significant representative values for entire masses of data which will stand for the essential meaning of the detailed facts. The average is such a value.

THE CONCEPT OF THE AVERAGE AS EMPLOYED IN EVERYDAY EXPERIENCE

The science of life insurance is based on the regularities observed in large numbers of cases. The records of mortality show the deaths of each age period per one thousand exposed at that age. It is not known when a particular person will die, but it is possible from a careful study of the mortality of a large number of persons to calculate the average number of years which a person of any given age will most likely live. This is called the expectation of life at that age and varies with the age of the insured.

In recent years the principle of adjusting wages to the requirements of a standard of living has been widely adopted. How shall we arrive at a satisfactory measurement of this standard? Certainly the experience of each separate family could not be made the basis of wage adjustment for the head of that family. It has been necessary to gather the detailed records of expenditures of many families covering food, rent, clothing, fuel and light, and sundries. From these budget records typical or average expenditures for food and each of the other items can be calculated for a family of a given size and social status.

At the milk station where babies are weighed the nurse has a chart showing the weight of a normal healthy child during each week of the first two years of life. By reference to this chart it is evident at once whether a baby is normal and if not how much its weight varies above or below normal. What is meant by normal? Normal weight means the average for a large number of healthy children of that particular age. We cannot speak of under- or overweight without assuming the concept of the average. Otherwise, the terms have no meaning.

Labor unions in a number of States, for example Massachusetts and New York, report regularly the proportion of their members out of work on a certain day each month. From these records, if accurate, we may learn the seasonal movement of unemployment in organized trades and how any month compares with the same month of the preceding year. For example, is employment in the building trades normal during the present month or year? In the year 1906 there was little unemployment and in the year 1908 large numbers were out of work. This extreme condition of unemployment was repeated again in 1914 at the opening of the World War. It will not be possible to judge whether or not employment is normal until we establish a level or trend over several months or years with which to compare present conditions.

Individual measurements are not usually significant except in relation to other individual measurements or more often to some typical value which *stands for a number of such measurements,* for example, average wage, average cost of living, the usual price, the general death-rate, the usual amount of unemployment at a particular time of year.

Business men particularly are interested in projecting past experience into the future; in predicting the requirements and prices for labor and materials and the market demand for products. All business has a large element of uncertainty. That which has taken place in the past forms a basis of judgment as to what will probably take place in the future and the past is a matter of record, or should be. This is the basis of business research. Conservative business men do not formulate policies on the basis of records of the past month or year alone. One month or year may be unusual. It is the combination and comparison of data over a period of the past which eliminates the unusual and establishes the norms or trend which form the basis of sound judgment and forecasting. Furthermore, any unusual event — a war, an earthquake, or a series of bank failures — will nullify the most careful estimates.

THE NATURE OF AN AVERAGE

An average is a representative value, actual or calculated, which, for a specific purpose, stands in the place of numerous individual measurements or estimates. It satisfies the desire for concentrated information. For the moment it disregards the variations among the items of the series and levels out all the differences. An average frees the mind of details and enables it to grasp the significance of the entire series. An average describes a series of varying individual values. *Above everything else, it should be a guiding principle not to accept any average without reference to the detailed measurements from which it is derived and of which it is presumed to be the best possible representative.*

The general functions of an average. Putting aside for the moment the fact that there are several kinds of averages, arrived at by different procedures, and having special uses, depending upon the purpose of the investigator and the character of the data with which he works, we wish to emphasize and illustrate the following general functions of an average.

I. A summation of values. Comparison of a number of series is rendered very simple by the use of average values which represent the detailed data. For example, the wages received by different classes of railway employees, when presented in frequency tables, would be very difficult to compare because of the size and complexity of the series. If, however, each series is or can be represented by an average value, the problem is greatly simplified. In a time series the fluctuation of individual yearly values often obscures the general movement, as of trade or immigration. If, however, averages for periods longer than a year, perhaps for five- or ten-year periods, are compared, the short time fluctuations are eliminated and the essential underlying movement is revealed.

A warning in comparing averages. *An average may be too simple.* Differences which are fundamental for correct interpretation may be covered up in a single expression. For example, the average price of a commodity for the year does not show monthly changes which may be of the greatest significance. The following illustration is taken from the volume on *Mines and Quarries* of the Twelfth Census.

TABLE 9. AVERAGE ADULT MALE EMPLOYEES BY MONTHS AND FOR THE YEAR, ANTHRACITE COAL INDUSTRY — 1902

January	110,018	July	6,493
February	110,760	August	7,610
March	109,165	September	8,136
April	109,190	October	34,773
May	53,169	November	105,516
June	16,301	December	110,393

Average for year — 65,127

The average for the year, 65,127, is not typical either of the full employment months of January through April and December, or of the slack months. In fact, the months of midsummer represent an entirely different employment situation because of the labor difficulties of that year. The average for the year does not represent any of the individual items from which it was derived.

The importance of homogeneity. In Chapter II illustrations were cited to show errors in conclusions arising from comparison of quantities which are not comparable. Averages may not be comparable because they have been derived from data representing widely different conditions and groups, not similar enough to admit of representation by a single expression — data which are not homogeneous. Let us suppose, for instance, that the wage data of one factory include the salaried office force while the facts from the payrolls of another similar factory do not. The arithmetic average wage of the first factory cannot be compared with that of the second, because the two quantities do not represent the same wage situation in both.

The arithmetic average of a series of wage data where wages of both men and women are included is not typical of either men's or women's wages. It is too low to represent the one and too high for the other. Moreover, suppose we desire to compare the average wages derived from two series, each of which includes both men and women but in different proportions. In Series A the average wage of the men is $20 per week and of the women $10. The number of men and women is equal. The average wage of the entire Series A, including both men and women, is $15. In Series B the average wage for men is $18 per week and for women $8, but there are three times as many men as women. The average for the entire Series B is $15.50,[1] which is higher than that for Series A, in spite of the obvious fact that the wages for both men and women in the second series are lower than in the first. It is clear that the two averages are not fair indices of the actual wage conditions. *An average, to be useful, must be typical of actual conditions, not merely a result of mathematical calculation.*

II. A type value from which to measure variability. The average may be calculated in order to summarize detailed data so as to free the mind from the burden of many items and to facilitate comparison in the manner just described. A no less important purpose may be to establish a point of departure from which to measure the variability of the individual cases.

Statistics is concerned with mass phenomena, but the individual cases are not to be disregarded. The knowledge that the weight of one school child differs from that of a second child has a very limited significance. Whether the difference in weight between A and B is more significant than that between B and C is difficult to judge. It is far more useful to know how the weight of each child is related to a typical weight for children of the given age and height; how much it is above or below that typical weight.

Having established a value representative of the entire group of phenomena, we approach the individual measurement with new interest. Our attention is no longer fixed on the difference of one individual measurement from another, but upon the difference of each individual value from the central value or average. This difference will be called *deviation from the average*. Is unemployment greater or less than usual at this time of the year? Are the workers' earnings high or low and to what extent? Is the output of the coal mines normal for the month of November? It is impossible to answer these questions without first having established some norm from which a conclusion may be reached as to the significance of individual values.

[1] This average is a weighted mean, the principle of which is explained later in the chapter. It is computed as follows:

$$\frac{(\$18 \times 3) + (\$8 \times 1)}{4} = \$15.50$$

Grouping about a central value. Not only is a single value in the series judged with reference to the central value or average, but likewise the variation of all the items of the series taken together is judged with reference to the average. This gives a picture of the grouping of the entire series about the average, which has been illustrated in graphic form in Chapter V. For example, it is possible to find out whether wages in organized trades tend to group more closely around the average (that is, to vary less from the average) than in unorganized trades. The statement has often been made that the labor union tends to level out wages and, by making wages more nearly alike, tends to restrict individual initiative. The truth or falsity of this statement may be tested by a comparison of the deviations of the wages of organized wage-earners from their average wage with similar deviations of the wages of unorganized wage-earners. *The degree of likeness or homogeneity* [1] *of the data is measured in this manner.* The less the total amount of the deviations, the greater is the homogeneity of the data, the closer is the grouping about the average, and the less is the variability among the individual items.

It is clear, therefore, that averages may be computed to serve as a point of departure, not only for judging a single value, but also for investigating the grouping of items in relation to the central value. This is important in deciding whether or not the average is typical. Averages do not have the same validity when computed from different kinds of data.

THE KINDS OF AVERAGE

We have described in some detail the need for an average, the nature of this measure, and the purposes to be accomplished by a descriptive value which characterizes an entire series. This may be arrived at in various ways. Therefore, more than one kind of average is possible. We shall discuss only those widely used in statistical work. It should be noted that the term average is used in the preceding discussion in a generic sense to include all forms. The chief forms of average are:

1. Mean or arithmetic average, both the simple and the weighted.
2. Median.
3. Geometric average.
4. Mode.

[1] Homogeneity and heterogeneity in *qualitative* series have been illustrated by the mingling of the wage data for *men and women.* In *quantitative* data homogeneity and heterogeneity refer to the degree of likeness or difference in the values of some particular characteristic, as the wages of *men* alone. The methods of measuring these for quantitative series will be discussed in Chapter IX.

The peculiar characteristics of each of these averages, how they summarize the items of the series, how they are arrived at, and some of their applications are set forth in the following pages. Why are there so many different ways of determining an average? The discussion will make clear why more than one kind of average is necessary.

THE ARITHMETIC AVERAGE — THE MEAN[1]

The *mean* is the most widely known and used of all averages. Every item of the given series is included in its computation. However, it may be computed from a knowledge of the total value of all the items and their number by simply dividing the one by the other without exact knowledge about any single item. For example, the average earnings may be computed from a knowledge of the total payroll and the number of men employed. A change in the value of any single item affects the mean. A very large or a very small value may seriously affect it. This is not true of the mode and the median which are values chosen to represent the series because of their positions in it. The mean denotes the size which the individual measurements would be if all were made exactly alike, while the total remained the same. The median and mode can be found only in series where items have been arranged according to magnitude, but the mean does not require any definite order of the items for its computation.

Computation from ungrouped data. To compute the value of the mean when the series consists of the original individual measurements of the weights of 1000 college freshmen, it is only required to sum all the values and divide by their number. The total of these individual values equals 134,408 pounds, and dividing by 1000 we find 134.41 pounds to be the *true average weight*. If we generalize this procedure for ungrouped measurements, in terms of symbols, we may let M represent the mean,[2] X the value of an individual measurement, N the number

[1] For definition and illustration of the Harmonic Mean, see G. U. Yule: *An Introduction to the Theory of Statistics*, 6th ed., 1922, pp. 128–29.

[2] An original magnitude will be designated X when dealing with a single series. When a second series is introduced, as in correlation, Y will be used to represent its values. The mid-value of a class-interval will be designated by m.

In explaining statistical processes we shall first direct attention to the procedure and its meaning and then use the formula or symbol merely to summarize the procedure in convenient form. In other words, the attempt will be made to present the explanation in non-technical form. The appeal will be made to common-sense and the logical faculties. The formula is too apt to be used in a blind quest after results without the student being able to explain what the results signify. The difficult task of the statistician is to decide if a particular procedure applies to the specific kind of data he is handling and for the purpose he has in mind. Correct calculations according to certain formulas do not necessarily bring forth valuable results. Formulas, when regarded with too great confidence, tend to become ends in themselves instead of means. This use of mathematical procedure tends to inhibit of items, and Σ the sum of the various items. The average may be expressed in terms of these symbols.

$$M = \frac{\Sigma X}{N}$$

— sats

Computation of the mean from grouped data. Very frequently, however, the procedure in computing the mean is not so simple as indicated in the above example, because the series does not show individual measurements in their original form, but the number of items between certain limits of value, called *class-intervals*. This arrangement in group form is a step in the condensation of data for greater convenience of handling, described and illustrated in Chapter V. When once the individual measurements have been grouped within certain class-intervals, the series does not show how the items are distributed between the limits of a single class. In order to compute the arithmetic mean from such a series, an assumption must be made concerning the distribution of items over class-intervals. *Uniform distribution over the class-interval is usually assumed*, and, therefore, in computing the mean, the mid-value (m) of each class is taken as the average value for all items of that class. The procedure is illustrated in the problem of computing the average weight of college Freshmen in Table 10.

The difference in the computation of the mean from grouped as compared with ungrouped data consists in the assumption of even distribution made necessary by the process of grouping. Instead of adding the original values between 90 and 95 pounds, we assign to each of the six items within that class the mid-value, 92.5 pounds, and multiply this mid-value by the number of items, six, in order to arrive at the total weight represented by this class, 555 pounds. This is done in succession for each class and the products when added give the total weight of the thousand individuals, 134,445 pounds, as compared with the sum of the actual measurements, ungrouped, which was 134,408 pounds. It will be noted that there is a difference between these totals of 37 pounds, which is insignificant in its effect upon the average, and is caused by our assumption of the mid-values of each of the class-intervals as the value for every item occurring within the given class. The advantages of using grouped data have been suggested in Chapter V, and will become more evident as we proceed in our discussion and analysis.

The effect of grouping upon the computation of the mean may prove

the usual checks of consistency and common-sense which are the most valuable possessions of the worker with quantitative data. Symbols should be regarded as a kind of shorthand, for the convenience of the reader and the user.

TABLE 10. COMPUTATION OF THE MEAN FROM GROUPED DATA — LONG METHOD

CLASS LIMITS [a] (pounds) (1)	MID-VALUE (m) (2)	FREQUENCY (f) (3)	COLUMN (2) TIMES (3) (mf) (4)
90– 94.9	92.5	6	555.0
95– 99.9	97.5	7	682.5
100–104.9	102.5	10	1,025.0
105–109.9	107.5	18	1,935.0
110–114.9	112.5	65	7,312.5
115–119.9	117.5	81	9,517.5
120–124.9	122.5	111	13,597.5
125–129.9	127.5	134	17,085.0
130–134.9	132.5	125	16,562.5
135–139.9	137.5	117	16,087.5
140–144.9	142.5	85	12,112.5
145–149.9	147.5	75	11,062.5
150–154.9	152.5	54	8,235.0
155–159.9	157.5	35	5,512.5
160–164.9	162.5	25	4,062.5
165–169.9	167.5	21	3,517.5
170–174.9	172.5	13	2,242.5
175–179.9	177.5	5	887.5
180–184.9	182.5	5	912.5
185–189.9	187.5	4	750.0
190–194.9	192.5	2	385.0
195–199.9	197.5	1	197.5
200–204.9	202.5	0	
205–209.9	207.5	1	207.5
		1000	134,445.0 pounds

$M = \frac{\Sigma mf}{N} = \frac{134{,}445}{1{,}000} = 134.445$ pounds, the mean weight from the grouped data, in five-pound classes. In the formula here employed, m represents the mid-value of each class-interval and f the number of cases in each class.

[a] This form of stating the upper limit of the class means that all values up to but not inclusive of 95 are grouped in the first interval, and is used instead of the form 90 and under 95. The original data are given *to the nearest tenth of a pound*. The class limits as stated above involve a slight error of $\frac{1}{20}$ of a pound because the original measurements at 90.0 pounds really range from $89\frac{19}{20}$ to $90\frac{1}{20}$, having been measured to the nearest tenth. The mid-value would be $92\frac{9}{20}$ instead of 92.5. To use these fractions in this problem is an *unnecessary refinement.*

to be a matter of serious importance, as another example illustrates. The frequency table employed in Chapter V, which classifies five hundred marks in English, indicates that the assumption of even distribution over the class-interval of five per cent is not in accord with the facts. There is decided concentration at certain values. The fixing of the limits of the class becomes important in the accuracy of the computation of the mean grade of the entire series. The points of concentration must first be determined experimentally from the original individual values, and then, if possible, these points should be located in the middle of the class-intervals adopted for final grouping. Two methods of grouping were set forth in Chapter V, page 78, to which the reader should refer. The effect of these two methods on the computation of the mean is shown in Table 11.

TABLE 11. COMPUTATION OF THE MEAN GRADE FROM DIFFERENT GROUPINGS

Grouping A				Grouping B			
Class Limits [a] (per cent) (1)	Mid-Value (m) (2)	Frequency (f) (3)	Column (2) Times (3) (mf) (4)	Class Limits (per cent) (1)	Mid-Value (m) (2)	Frequency (f) (3)	Column (2) Times (3) (mf) (4)
19.5 less than 24.5	22	22	484	17.5 less than 22.5	20	21	420
24.5 " " 29.5	27	21	567	22.5 " " 27.5	25	21	525
29.5 " " 34.5	32	47	1504	27.5 " " 32.5	30	42	1260
34.5 " " 39.5	37	59	2183	32.5 " " 37.5	35	54	1890
39.5 " " 44.5	42	61	2562	37.5 " " 42.5	40	68	2720
44.5 " " 49.5	47	78	3666	42.5 " " 47.5	45	64	2880
49.5 " " 54.5	52	66	3432	47.5 " " 52.5	50	78	3900
54.5 " " 59.5	57	25	1425	52.5 " " 57.5	55	27	1485
59.5 " " 64.5	62	51	3162	57.5 " " 62.5	60	45	2700
64.5 " " 69.5	67	23	1541	62.5 " " 67.5	65	27	1755
69.5 " " 74.5	72	22	1584	67.5 " " 72.5	70	28	1960
74.5 " " 79.5	77	12	924	72.5 " " 77.5	75	11	825
79.5 " " 84.5	82	7	574	77.5 " " 82.5	80	8	640
84.5 " " 89.5	87	3	261	82.5 " " 87.5	85	3	255
89.5 " " 94.5	92	3	276	87.5 " " 92.5	90	3	270
		500	24,145			500	234,85

$$M = \frac{\Sigma\, mf}{N} = \frac{24{,}145}{500} = 48.3 \text{ per cent} \qquad M = \frac{23{,}485}{500} = 47.0 \text{ per cent}$$

[a] The original marks are given to the nearest per cent. Therefore, the lower limit of the first interval is 19.5 and the class extends up to but not inclusive of 24.5. "Less than" is another method of defining the upper limit of the class.

THE WEIGHTED ARITHMETIC AVERAGE

/ An illustration will make evident a limitation involved in the use of the method of simple averaging. The mean height of three children, whose heights are 31, 32, and 33 inches respectively, is equal to

$$\frac{31 + 32 + 33}{3} = 32 \text{ inches.}$$

Likewise, of seven other children whose heights are 35, 36, 37, 38, 39, 40, and 41 inches, the mean height is equal to

$$\frac{35+36+37+38+39+40+41}{7} = 38 \text{ inches.}$$

But the mean height of these two groups, when combined, is not

$$\frac{32+38}{2} = 35 \text{ inches,}$$

as will be seen by combining all the individual heights,

$$\frac{31+32+33+35+36+37+38+39+40+41}{10} = 36.2 \text{ inches.}$$

To take the simple mean of the two averages is not sound, because a different number of items makes up the two groups. The second group includes over twice as many as are found in the first, and yet, by simple averaging of the two averages, each group is given equal importance in the result. If we assign to the two groups relative weights in proportion to the number of items in each, 3 and 7, we may average the two averages with accuracy as follows: $\frac{(32\times 3)+(38\times 7)}{10} = 36.2$ inches. In averaging, it will often be found necessary to give unequal importance to the quantities combined. This is done by multiplying the original magnitudes by weights which express their relative importance, summing up the products, and dividing by the sum of the weights. In terms of symbols, the weighted mean is equal to

$$\frac{\Sigma XW}{\Sigma W}$$

(In addition to the symbols previously explained, W represents the numerical weights.)

The weighted mean is not an independent kind of average. There is no difference in the fundamental principle of its computation from that of the simple mean, and the distinction between the two is sometimes presented as a formal rather than an essential difference. The series of values is first changed in its formation by the introduction of a new set of quantities, with the object of giving to the items of the original series, which are being combined in the average, an influence in proportion to their importance as indicated by the relative weights employed.

The question arises whether we have not already used the weighted mean in this chapter. Are not the frequencies to be regarded as weights, according to the explanation given? These frequencies, without doubt, do indicate the relative importance of each mid-point value in the series of class-intervals. There is no error in regarding the frequencies as weights, provided it is recognized that a simple mean cannot be computed for a frequency distribution without following exactly the same procedure as for the weighted mean. In other words, the distinction between the simple and weighted mean becomes a mere matter of the form of the series, depending on whether the items are combined into an average from the original ungrouped measurements, or have been concentrated, first, in the form of a frequency distribution. *This is neither a clear-cut nor a useful distinction.* There is no real reason why values of the type computed in the early part of this chapter should be regarded as weighted means. Since the frequency table is made up from individual measurements, the obvious and only way of computing a mean, employing all the individual items, is to multiply the mid-values of the classes by the number of cases having these values and to add the products. But this is practically equivalent to putting all the individual values in a series and adding them. The frequency table is an expedient to avoid the repetition of closely similar values, or the stating of individual values occurring within a certain range or class-interval. The series is concentrated by stating how many cases occur at each given value, or within each value class. There is no choice as to whether the frequencies shall be used or not. All observations must be included. On the adding machine the procedure of summation is no different, whether the data are left in the original form of individual values or are grouped by the frequency table. All such computations should be termed *simple means*. The essential in this mean is the summation of individual items, as originally measured or estimated, or as grouped for convenience in handling. The device of grouping in the frequency table is merely a step in the process of summation.

But this does not close our discussion of the weighted mean. The problem of choosing and applying weights in averaging is a real one. For example, it is desired to compute a mean retail price of a commodity which is being sold in varying quantities at different prices, 12, 14, 18, and 20 cents per pound. The facts are for a single city and the prices are averages obtained by the investigation of a limited number of stores of different types in various sections of the city. A simple mean would be equal to $\frac{12 + 14 + 18 + 20}{4} = 16$ cents per pound. A weighted mean changes the series of prices before combining them in a manner to give proper importance to the different prices by introducing as weights the amount sold at each price. Since the investigation covered only part of the stores in any section of the city the data used for weights could be secured only by special supplementary investigation or by estimates. The result appears as follows:

PRICE (cents) X (1)	WEIGHTS ESTIMATED QUANTITY SOLD (pounds) W (2)	XW (3)
12	100,000	$12,000
14	60,000	8,400
18	50,000	9,000
20	20,000	4,000
	230,000	33,400

$$M = \frac{\Sigma XW}{\Sigma W} = \frac{33,400}{230,000} = 14.5 \text{ cents per pound}$$

This method of averaging reduces the mean price from 16 cents to 14.5 cents per pound, because of the larger sales at the lower price and the greater influence given to this price by the system of weights in column (2).

That the weighted mean correctly describes the transactions as a whole, is shown by multiplying it by the total units sold which equals the total value of sales. On the other hand, if we multiply the simple mean by the total units sold, the calculated amount is considerably larger than the actual total value of sales.

If we do not use weights in this problem, it is assumed that about the same amounts of the commodity changed hands at each price, which we know is not true. The investigation which secured the price data did not record the facts necessary for weighting because only a limited number of stores were visited. These quantities had to be collected or estimated by supplementary investigation. This is a very different situation from that where frequencies are used in computing the mean. In the latter case the frequencies are the numbers of original measurements, and there is no choice as to their use. In weighting it is often necessary to estimate the quantities used more or less arbitrarily. Usually some logical basis can be found for choosing the values to indicate the relative importance of the items combined.

Combining relative quantities. The reader is requested to turn to

Chapter II and review the criticism of the Aldrich Wage Report, covering wage changes from 1860 to 1891 in the brewing industry, which was used as an example of doubtful statistical method. The purpose was to summarize the wage changes in the various subdivisions of the brewing industry — a single figure for the entire industry. A simple mean was employed which gave equal weight to each separate percentage in spite of the fact that there was only one master brewer, with a large wage increase, and many common laborers, whose percentage increase was much less. Here the logical procedure would have been to weight each percentage of increase by the number of workers affected by such change. However, this would have involved using as weights the total workers in the industry. The investigation covered only one establishment in the industry, a very small sample. To secure the proper weights for the separate percentages in this and other industries represented in the report would have required supplementary investigations. However, it is clear that logical procedure in this situation requires that the quantities averaged shall first be adjusted to express accurately their relative importance. Weighting is the device by which this may be done.

Another example of the application of weights in combining relatives will be found in Chapter II, page 16. During the period under consideration the prices of various items of the family budget, food, rent, clothing, fuel and light, and sundries increased at different rates. A combined percentage change in the cost of the entire family budget was desired, in order that the wages of labor could be adjusted in accordance with the increase in the cost of living. A simple mean of the five percentage changes was not logical because this procedure would assign equal importance to all the items of the family budget, whereas some items absorb much more of the family income than others. Clearly, these percentage changes in prices, when combined, should be weighted according to the percentages of the total family expenditures which are apportioned, on the average, to food, rent, and other items. Other than price investigation is required to ascertain these proportions. A considerable number of family accounts or budgets must be collected and analyzed, the items must be tabulated and averaged, and the proportion of each item to the total family budget must be computed. These proportions may then be used as weights in averaging the price changes.

The matter of combining relatives will receive further attention in Chapter X on "Index Numbers."

Combining averages. One final illustration may be given of the need for weights, in combining averages. Let us suppose that investigation of a specific branch of an industry in ten cities has secured an average wage for each city. There are significant differences in the averages, and the number of workers in this occupation varies widely in the different cities. Our purpose is to secure a single figure to express the average wage in this branch of industry for the ten cities combined. The investigation has covered only a sample of representative establishments in each city. Therefore, it does not include wage data for all the workers, although the average wage is assumed to be typical in that branch of industry for each particular city.

This is a problem requiring the averaging of averages. A simple mean gives equal importance to each city, regardless of the different numbers of workers affected by the various wage rates. The average wage for each city must be weighted by the number of workers affected in that city. These weights can be ascertained from a manufacturing census, or they must be estimated. *If we would avoid error great caution must be exercised in combining either relatives, rates or averages.*

SHORT METHOD OF COMPUTING THE MEAN

That the algebraic sum of the deviations from the mean must always equal zero[1] *is a fundamental characteristic.* By *deviation* we mean the difference between any individual value in the series and the average of that series, those deviations below the average being designated *minus* and those above the average *plus.* *In other words, the sum of the positive deviations from the mean must exactly equal the sum of the negative deviations.* This must be true because of our definition of the mean as the value which would result if all individual values were made the same, the total remaining constant. Any excess of an individual value above the mean, according to this definition, would have to be balanced in amount by a corresponding amount below the mean. Let us examine a simple illustration where the distribution about the mean is perfectly symmetrical.

The "x" in Table 12 represents the difference between any specific value and the average, 12 years, called the *deviation* of the value from the average. The sign may be plus or minus, depending upon whether the value is more or less than the average, as 10 years -12 years $= -2$ in

[1] Let the observations be $X_1, X_2 \dots X_N$ and their mean value M. Then, according to the definition of the mean, $\frac{X_1 + X_2 + \dots X^N}{N} = \text{M}$, and $X_1 + X_2 + \dots X_N = NM$. Transposing NM to the left-hand side of the equation, we have

$$(X_1 - M) + (X_2 - M) + \dots (X_N - M) = 0.$$

This equation is the sum of the deviations of the items from the average and equals zero.

TABLE 12. THE ALGEBRAIC SUM OF DEVIATIONS FROM THE MEAN AGE TO THE NEAREST BIRTHDAY

AGE (years) m (1)	f (2)	mf (3)	x (4)	fx (5)
10	1	10	−2	−2
11	3	33	−1	−3
12	4	48	0	0
13	3	39	+1	+3
14	1	14	+2	+2
	12	144		0

$$M=\frac{\Sigma mf}{N}=\frac{144}{12}=12 \text{ years.}$$

column (4).[1] The products of the deviations times the frequencies are given in column (5). The algebraic sum of column (5) equals zero, the sum of all the negative and positive deviations being equal.

Frequency distributions are not usually so regular or symmetrical in their groupings about the mean. Nevertheless, this characteristic of the mean remains true for other than symmetrical distributions, as will appear from Table 13.

TABLE 13. THE ALGEBRAIC SUM OF DEVIATIONS FROM THE MEAN AGE TO THE NEAREST BIRTHDAY

AGE (years) m (1)	f (2)	mf (3)	x (4)	$-fx$ (5)	$+fx$ (6)
5	4	20	−3.516	−14.064	
6	9	54	−2.516	−22.644	
7	50	350	−1.516	−75.800	
8	86	688	− .516	−44.376	
9	54	486	+ .484		+26.136
10	24	240	+1.484		+35.616
11	13	143	+2.484		+32.292
12	10	120	+3.484		+34.840
13	5	65	+4.484		+22.420
14	1	14	+5.484		+ 5.484
	256	2180		−156.884	+156.788

$$M=\frac{\Sigma mf}{N}=\frac{2180}{256}=8.516 \text{ years.}$$

[1] The mean must always be subtracted from the magnitude to obtain the deviation with proper sign. ($m - M = x$.)

The mean is calculated to three decimals merely to demonstrate that the sum of all the positive and negative deviations are equal, when measured from the mean. The student should try the effect of using one decimal, 8.5 years, and two decimals, 8.52 years. As more decimals are used, the sums of the positive and negative values in columns (5) and (6) approach equality. It must be remembered that the deviation of the mid-value of the class must always be multiplied by the frequency of that class. *It is not the algebraic sum of column* (4) *but of columns* (5) *and* (6) *that must equal zero.* This could be demonstrated from any of the frequency tables employed in the preceding pages.

Short method. The short method of computing the mean is based upon the characteristic that the algebraic sum of the deviations must equal zero. To illustrate this method let us use the frequency distribution of the weights of college Freshmen, arranged in five-pound intervals, Table 14, page 98, the mean weight of which we have already computed by the long method, Table 10, page 89.

First, by inspection, and without knowing the value of the true mean, we select a value in the *middle of a class-interval* and call it the *guessed average.* Suppose we estimate 132.5 pounds as the guessed average. In this class there are 125 men whose weights range from 130 to 135 pounds, and whose average weight is assumed to be 132.5 pounds. We measure the negative or positive deviation of the mid-value of each class-interval from this guessed average. If our guess proves to be the true mean, which does not usually happen, we know it at once by the fact that the positive and negative deviations, after being multiplied by their respective frequencies, are equal.

One special rule must be observed in the short method. *The deviations from the guessed average as in column* (3) *must be stated in steps or intervals* rather than in actual units of measurement, in this case pounds. This rule must be followed whatever the size of the *uniform* class-interval may be, as, 5 pounds, 1 year, 50 cents. This procedure avoids fractions in the computation. The step deviations in column (3) are multiplied by the frequencies and the negative and positive products are recorded separately in columns (4) and (5). When columns (4) and (5) are each totaled the positive deviations exceed the negative by 389 step-units, which indicates that the guessed average was estimated at a value below the true mean, thus causing the total positive deviations to be greater than the negative.

To make the guessed average the true mean it is necessary to adjust it so that the positive and negative deviations will become equal. This is done by computing from the table a correction factor (c) by dividing the algebraic

TABLE 14. SHORT METHOD OF COMPUTING THE MEAN

WEIGHT (pounds) (1)	f (2)	d[a] (steps or intervals) (3)	$-fd$ (4)	$+fd$ (5)
90– 94.9	6	−8	-48	
95– 99.9	7	−7	-49	
100–104.9	10	−6	-60	
105–109.9	18	−5	-90	
110–114.9	65	−4	-260	
115–119.9	81	−3	-243	
120–124.9	111	−2	-222	
125–129.9	134	−1	-134	
130–134.9	$\boxed{125}$ 500th →	0		
135–139.9 137½	117	+1		117
140–144.9	85	+2		170
145–149.9	75	+3		225
150–154.9	54	+4		216
155–159.9	35	+5		175
160–164.9	25	+6		150
165–169.9	21	+7		147
170–174.9	13	+8		104
175–179.9	5	+9		45
180–184.9	5	+10		50
185–189.9	4	+11		44
190–194.9	2	+12		24
195–199.9	1	+13		13
200–204.9	0	+14		0
205–209.9	1	+15		15
	1000		−1106	+1495

$$C = \frac{\Sigma fd}{N} = \frac{+\,1495 - 1106}{1000} = +.389 \text{ steps or intervals.}$$

$$= +.389 \text{ steps} \times 5 \text{ (interval) pounds} = +1.945 \text{ pounds.}$$

$$G.\,A. + C = 132.5 + 1.945 = 134.445 \text{ pounds} = \text{True mean.}$$

[a] In the preceding tables x was used to denote the deviation in terms of the original unit of measurement. In this table d is used in order to indicate that the deviation is in *steps* and that we are using the short method. This is done in all succeeding tables.

summation of columns (4) and (5) by the number of cases, $\frac{\Sigma fd}{N}$, which equals $\frac{+389}{1000} = +\,.389$ of a *step*. Since the correction factor, + .389, is in terms of steps and not in pounds, we cannot add it to the guessed average until we have reduced it to pounds by multiplying by the size of the class-interval, five pounds (+ .389 × 5 pounds = +1.945 pounds). The guessed average may be adjusted by adding 1.945 pounds and the true mean is the result, 134.445 pounds.

Formula. True Mean = $G.A. + c$, in which $G.A.$ is a guessed average taken at the mid-value of any class-interval and c is a correction factor to be added to the guessed average. Care must be taken to preserve algebraic signs. The correction factor, c, is computed from the table according to the formula, $c = \frac{\Sigma fd}{N}$, the summation being algebraic. Therefore,

$$\text{Mean} = G.A. + \frac{\Sigma fd}{N} = 132.5 \text{ pounds} + (+ .389 \times 5 \text{ pounds}) = 134.445 \text{ pounds}.$$

The accuracy of this result may be readily checked by choosing another guessed average at the mid-value of some other interval and carrying through similar computations. Precisely the same true mean should be obtained, if the original computation was correct. The student is advised to choose several different guessed averages in the table and test for himself the validity of this short method. The caution is repeated not to forget to reduce the correction factor (c), which is always in terms of steps, to the units which characterize the actual class-interval before using c for adjusting the guessed average.

Since most distributions do not exceed twenty class-intervals, the calculations will be simple, if the guessed average is located near the middle of the distribution, because the steps need not exceed ten on either side of the guessed average. In case the number of intervals is not large enough to make this consideration important, it is best to locate the guessed average in the class where there is greatest concentration of frequencies.

The mean is the same to the second decimal with ten-pound intervals, just half the number of classes, as with the five-pound intervals. The wider grouping, with consequent fewer classes, is simpler to manipulate, and the difference in the mean is insignificant. Matters of this kind should be determined experimentally for a given type of data. Moreover, in Table 15 the sum of all the negative deviations exceeds the positive by 55 unit steps, which indicates that the average was guessed above the true mean. This situation is opposite the one illustrated in Table 14. The correction factor is minus and must be subtracted from the guessed average, 135 pounds, after reducing the step-units to pounds. *The student should note that in this case the class-interval is 10 pounds and, therefore, the −.055 steps is multiplied by ten.*

It should be emphasized that the short method yields the same result as the long method used in Table 10. The short method is not one of approximation but an accurate mathematical calculation.

55229

TABLE 15. CALCULATION OF THE MEAN WEIGHT FROM TEN-POUND INTERVALS

WEIGHT (pounds) (1)	f (2)	d (steps) (3)	$-fd$ (4)	$+fd$ (5)
90– 99.9	13	−4	52	
100–109.9	28	−3	84	
110–119.9	146	−2	292	
120–129.9	245	−1	245	
130–139.9	242	0		
140–149.9	160	+1		160
150–159.9	89	+2		178
160–169.9	46	+3		138
170–179.9	18	+4		72
180–189.9	9	+5		45
190–199.9	3	+6		18
200–209.9	1	+7		7
	1000		−673	+618

The guessed average is taken at 135 pounds, the mid-value of the class which is distributed from 130 to 140 pounds.

$$c = \frac{\Sigma fd}{N} = \frac{-673 + 618}{1000} = \frac{-55}{1000} = -.055 \text{ steps.}$$

$-.055$ steps × 10 pounds $= -.55$ pounds.

135 pounds + $(-.55$ pounds$)$ = 134.45 pounds = True mean.

In Table 16, page 101, the application of the short method is illustrated in a case where the class-interval is a fraction, a half-dollar. These are the same data as were used in Figure 5, page 63.

Since the sum of all the positive deviations is larger than that of the negative deviations, the position of the guessed average is below the true mean. The correction in steps is reduced to cents by multiplying by .50, since the class-interval in this problem is a half-dollar.

The long method in this problem would have involved taking the products of the mid-values of the class-intervals, \$2.25, \$2.75, \$3.25, etc., by their respective frequencies, a far more laborious task.

When the class-intervals are already expressed as single units, as in the age problem on page 96, the correction factor is already in proper form for use in adjusting the guessed average. In this case the size of the class-interval, one year, and the size of the step are identical.

Short method when the intervals are not uniform. The short method is easily adapted to distributions having *any size of class-interval*, provided only the interval is *uniform in size throughout the entire range*. The examples used in the preceding pages have all been of this type. The

Table 16. Computation of Mean Earnings — Short Method

Classified Earnings (1)	*f* (2)	*d* (steps) (3)	−*fd* (4)	+*fd* (5)
$2.00–2.49	6	−4	24	
2.50–2.99	16	−3	48	
3.00–3.49	34	−2	68	
3.50–3.99	61	−1	61	
4.00–4.49	66	0		
4.50–4.99	57	+1		57
5.00–5.49	37	+2		74
5.50–5.99	28	+3		84
6.00–6.49	9	+4		36
6.50–6.99	6	+5		30
7.00–7.49	8	+6		48
7.50–7.99	8	+7		56
	336		−201	+385

G.A. = $4.25, the mid-value of the class $4.00 – 4.49.

$$c = \frac{\Sigma fd}{N} = \frac{+385 - 201}{336} = \frac{+184}{336} = +.548 \text{ steps.}$$

+.548 steps × .50 = +27 cents.

True mean = *G.A.* + *c* = $4.25 + .27 = $4.52.

method may also be used when the *intervals of the distribution are not uniform,* usually with greater difficulty and with less saving of time because fractional step-deviations are likely to be involved in the computation. The distributions in Table 17 are illustrative.

In Table 17 A the ages of workers are grouped in five-year intervals under 40 years and in ten-year intervals above 40 years. The last three intervals are twice the size of the others. In column (3) the step-deviations are all expressed in the form, 5 years = one step. The guessed average is 27.5 years. The mid-value of the interval 40–49, which marks the change to a ten-year grouping, is 45 years. The deviation from the guessed average, 27.5 years, is 45 − 27.5 = 17.5 years, which is 3.5 steps of five years each. The next two classes are each two steps higher in value, that is 5.5 and 7.5 respectively. The step-deviations of column (3) are fractional values which interfere more or less with the computation of $c = \frac{\Sigma fd}{N}$. In other respects, the method is the same as explained in former examples. Before adjusting the guessed average to obtain the true mean, *c in step-deviations must be multiplied by five years to reduce the correction to the unit of the problem.*

Table 17. Short Method — Unequal Class-Intervals

A — Ages of Workers			B — Ages of Workers More Detailed		
Class Limits (years) (1)	*f* (2)	*d* (steps) (3)	Class Limits (years) (1)	*f* (2)	*d* (steps) (3)
15–19	10	−2	15 and under 16	5	−2.4
20–24	50	−1	16 " " 17	10	−2.2
25–29	100	0	17 " " 18	15	−2.0
30–34	75	+1	18 " " 19	25	−1.8
35–39	40	+2	19 " " 20	40	−1.6
40–49	25	+3.5	20 " " 25	275	−1.0
50–59	10	+5.5	25 " " 30	300	0
60–69	5	+7.5	30 " " 35	400	+1.0
			35 " " 40	200	+2.0
			40 " " 50	100	+3.5
			50 " " 60	50	+5.5
	315			1420	
G.A. = 27.5 years.			*G.A.* = 27.5 years		

In Table 17 B three intervals of different sizes are used, one-year, five-year and ten-year groups. Again all the deviations are expressed in the form, five years = one step. The guessed average is 27.5 years. The values entered in column (3) are obtained by taking the deviation of the mid-value of each interval from 27.5 years, and dividing this deviation by five, to express each step in comparable units. In the one-year and ten-year intervals this procedure results in fractional step deviations, the former being one fifth of a step and the latter being two steps each. Again, the computation of *c* involves fractions but, otherwise, the method is the same as already explained.

It is desirable that the class-intervals of a frequency distribution should be uniform in size, or should be able to be made so by combination of intervals, in order to facilitate the use of the short method without involving fractions in the computations. If fractions are introduced it becomes doubtful whether time is saved and accuracy is promoted by the use of the short method. The decision depends upon whether hand-methods of computation are used, and how much fractional work is imposed by the irregular class-intervals of the particular distribution.

A final modification of the short method should be noted. This method expresses the deviations in the original unit of the problem, and, therefore, avoids the necessity of reducing the correction factor back to the original unit by multiplying by the size of the class-interval. For example, in Table 15 on page 100, where the guessed average is 135 pounds, the minus step-deviations would be stated −10 pounds, −20 pounds, −30 pounds, −40 pounds, and the plus deviations would be stated in a similar manner. In other respects, the computation of *c* is the same as shown in that table. When computed, *c* is already expressed in pounds and may be used to adjust the guessed average at once. This method should not be used in distributions where the class-interval is uniform in size because the step-deviation method is easier and quicker in computation.

CRITERIA FOR JUDGING AN AVERAGE

It may assist the student at this point to examine some of the criteria by which an average is judged. The various kinds of average differ in respect to the following:

1. The utilization in their computation of all or only part of the individual values.
2. The necessity of arranging the data in order of magnitude.
3. Special advantages from the point of view of arithmetical and algebraic treatment.
4. Effect of the size and the number of individual values.
5. Definiteness of their values.
6. The facility of computation.
7. The ease of comprehension by the reader.
8. The effect of extreme variants.

Limitations of the mean. Too great dependence upon a single value is dangerous. The average chosen should be typical of the actual conditions to be described. Any single value may prove misleading when used to describe a given series of quantitative data. *Approximately the same mean value may be computed from frequency distributions which are utterly different in their internal structure.* For example, let us compare the detailed wage classifications in two different establishments.

Examination of the frequency distributions and the corresponding graphic representation of the facts in Figure 10 indicates very different wage conditions in the two factories. However, the difference in the mean wage is only four cents. Of course, the mean wage of Factory II is not a representative value for this group. The two wage situations are not comparable by any such simple device as the mean wage, which does

Table 18. Comparison of Two Wage Distributions

Factory I		Factory II	
Wage	Number	Wage	Number
$10–10.99	6	8– 8.99	5
11–11.99	14	9– 9.99	12
12–12.99	22	10–10.99	25
13–13.99	45	11–11.99	44
14–14.99	80	12–12.99	36
15–15.99	60	13–13.99	18
16–16.99	35	14–14.99	12
17–17.99	15	15–15.99	15
18–18.99	10	16–16.99	22
19–19.99	5	17–17.99	36
Total....	292	18–18.99	40
		19–19.99	28
		20–20.99	6
Mean I = $14.79		21–21.99	1
Mean II = $14.83		Total....	300

not reveal the actual differences. The mean represents the wages of Factory I very well because there is a decided concentration at or near a

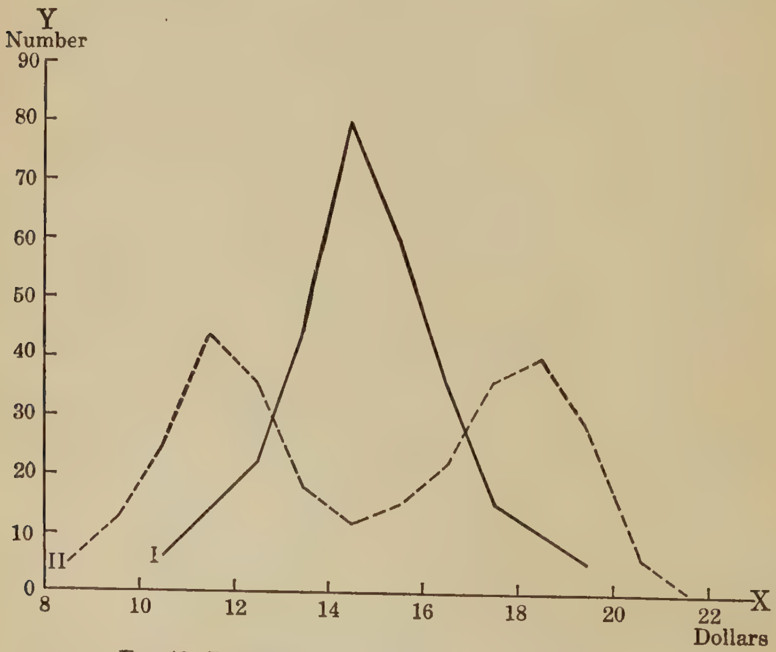


Fig. 10. Distribution of Wages in Two Factories

(Data from Table 18.)

central value as shown by the high point of the frequency polygon, and a fairly regular grouping of the data about this central value. In Factory II, however, the exact opposite is true. The mean falls at a value where there are the fewest wage items and there are two points of concentration, neither of which is represented by the mean. Possibly the wages of both men and women are included in this distribution, which would account for the two points of concentration. If this is true then no single value could represent the entire group. The utility of an average depends upon more than these arithmetical computations.

Effect of extreme variants upon the mean. The addition of a few measurements very high or very low in value may destroy the representative character of the mean. For example, the call money rate goes soaring for one or two critical days of the month. The mean rate of interest on call money for that month may be so affected by these extreme rates as not to represent the general situation prevailing throughout the rest of the month. A few very large incomes will produce a mean income far above that which is representative of the great mass of income receivers. One or two large checks among the contributions in the church collection conceal the usual or typical contribution, if the mean is used to describe the typical value.

It is evident that this form of average is greatly affected by the exceptional and the unusual. Under such circumstances, as suggested in the illustrations, the mean breaks down and some other way of arriving at a typical value must be devised. We shall examine another kind of average in the next chapter to discover, among other things, whether it is capable of meeting satisfactorily this defect.

SUMMARY

In the light of the criteria for judging an average, the mean may be characterized as possessing a special advantage from the point of arithmetical and algebraic treatment. In its determination the mean is based upon all the values, which need not be arranged in order of size for the computation, *but which should be so arranged that the reader may judge the significance and limitations of this form of average.* It is affected by both the size and the number of items, is rigidly defined, is generally understood, and, if the short method is used, is rather easily computed.

READINGS

(Most texts treat all the kinds of averages in the same chapter. We are concerned at present with the arithmetic average and its applications.)

Rugg, H. O., *Statistical Methods Applied to Education*, chap. 5.
Mills, F. C., *Statistical Methods Applied to Economics and Business*, chap. 4.

King, W. I., *Elements of Statistical Method*, chap. 12.

Yule, G. U., *An Introduction to the Theory of Statistics*, 6th ed., chap. 7.

Bowley, A. L., *Elements of Statistics*, 4th ed., part I, chap. 5.

Jerome, Harry, *Statistical Method*, chap. 7.

Secrist, Horace, *An Introduction to Statistical Methods*, chap. 8.

—— ——, *Readings and Problems in Statistical Methods*, chap. 7.

Zizek, Franz, *Statistical Averages*. Translated by Warren M. Persons, part I, chap. 6, and part II, chap. 2.

REFERENCES

Hoffman, F. L., *Insurance Science and Economics*, chap. 9. (Applications of averages.)

Mitchell, H. H., and Grindley, H. S., *The Element of Uncertainty in the Interpretation of Feeding Experiments*, Bulletin 165, Agricultural Experiment Station, University of Illinois, July, 1913, pp. 463–72. (An excellent presentation of the theory of averages and their application.)

Jones, D. C., *A First Course in Statistics*, chaps. 4 and 5.

Kelley, Truman L., *Statistical Method*, chap. 3.

Rietz, H. L. (Editor), *Handbook of Mathematical Statistics*, chap. 2, prepared by Professor Rietz.

Jerome, Harry, *Statistical Method*, chap. 8. (Professor Jerome discusses "Ratios and Coefficients" which are appropriately considered in connection with averages. For applications to Vital Statistics refer to G. C. Whipple, *Vital Statistics;* Raymond Pearl, *Medical Biometry and Statistics*, chap. 7; Arthur Newsholme, *The Elements of Vital Statistics*.)

[Details as to publisher and date of publication of the readings and references given at the close of each chapter are found in the alphabetical list in Appendix A.]

CHAPTER VII

METHODS OF SUMMARIZATION AND DESCRIPTION — THE MEDIAN AND THE GEOMETRIC MEAN

The median — a position average. It cannot be determined until the values are arranged in order of size. It is not influenced by extreme variants to the same extent as is the mean. Therefore, it proves a better form of average to use if we wish to avoid the influence of extreme variants. *However, the student is warned at the outset that the median average may not prove to be a representative value.* A case in point are the wage data of Factory II on page 104. In this distribution the median wage falls at a value where few of the wage items are located. Therefore, it could not fairly represent such a series.

The median may be defined as the *value* which is exceeded by one half of the measurements and of which one half of the measurements fall short. We locate it in the entire range of the distribution so that one half of the values fall above it and one half below it. *It represents the entire series, if at all, by virtue of its position.* It is clear that in any array of values which includes an odd number of items the median value is that of the middle case. When an even number of values is included in the array the median value is located between the two middle items. If it happens that these two middle values are identical then either may be chosen as the median value.

The effect of extreme variants upon mean and median. Let us observe the comparative results of these two methods of averaging, using a distribution of annual personal incomes under $4000 for the United States in 1918.[1] (Table 19.)

The method of computing the mean has been made clear. The median income, $1122, is that income in the entire distribution which has as many incomes below it as above it. The exact method of computation will be explained later. The largest number of income receivers is in the class $900 to 1000. Therefore, the income most often repeated in the entire array of individual incomes would be at about $950, the mid-value of this class. It will be observed that the mean income is over $300 more than this amount. The mean is influenced by the incomes far above the point of greatest concentration in the array, which pull it upward toward

[1] *The Income in the United States*, vol. I, pp. 132–33, National Bureau of Economic Research.

Table 19. Personal Incomes in the United States under $4000, 1918

Income in $100 Groups (1)	m (2)	f (thousands) (3)	Income in $100 Groups (1)	m (2)	f (thousands) (3)
$0– 100[a]	$50	63	2000–2100	2050	550
100– 200	150	104	2100–2200	2150	463
200– 300	250	209	2200–2300	2250	395
300– 400	350	490	2300–2400	2350	340
400– 500	450	962	2400–2500	2450	295
500– 600	550	1550	2500–2600	2550	259
600– 700	650	2154	2600–2700	2650	228
700– 800	750	2668	2700–2800	2750	201
800– 900	850	3013	2800–2900	2850	179
900–1000	950	3145	2900–3000	2950	154
1000–1100	1050	3074	3000–3100	3050	143
1100–1200	1150	2851	3100–3200	3150	128
1200–1300	1250	2535	3200–3300	3250	116
1300–1400	1350	2206	3300–3400	3350	105
1400–1500	1450	1832	3400–3500	3450	95
1500–1600	1550	1513	3500–3600	3550	86
1600–1700	1650	1234	3600–3700	3650	79
1700–1800	1750	1000	3700–3800	3750	73
1800–1900	1850	811	3800–3900	3850	67
1900–2000	1950	664	3900–4000	3950	62
Total number of incomes under $4000.					36,096

Mean income = $1255
Median income = $1122

[a] This statement of the limits of the class, 0 to 100, is used instead of the form 0 and under 100, or 0 and less than 100, or 0 to 99. It is not a clear definition of the upper limit of the class and should be avoided as a rule.

the higher values. In the year 1918 ten incomes were recorded over $4,000,000 each. If these had been included in the array the mean would have been still further influenced and would have proved still less representative of the entire distribution. *On the other hand, these ten incomes would have very little effect on the median, since only their number and not their extreme size influences the result.* It would be necessary to count only five more cases through the array to reach the new median value which would include the ten high incomes.

Another simple example indicates the relative effect of extreme variants upon the mean and the median. We wish to average the following eleven values: 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23. Since the number of items is odd, the median will be the mid-value, 13, with five items above and five below it. The mean is also 13, because of the perfectly symmetrical arrangement of the items around the central value. Now add two extreme values, 60 and 70, to the original number, making 13 items in all. The median will be 15 now, with six items on either side of it. The median value is changed only a small amount by the inclusion of the extreme variants, but how about the mean? It has been changed from 13 to 21, due to the addition of two extreme values *whose number and size both affect the result.*

METHODS OF DETERMINING THE MEDIAN VALUE

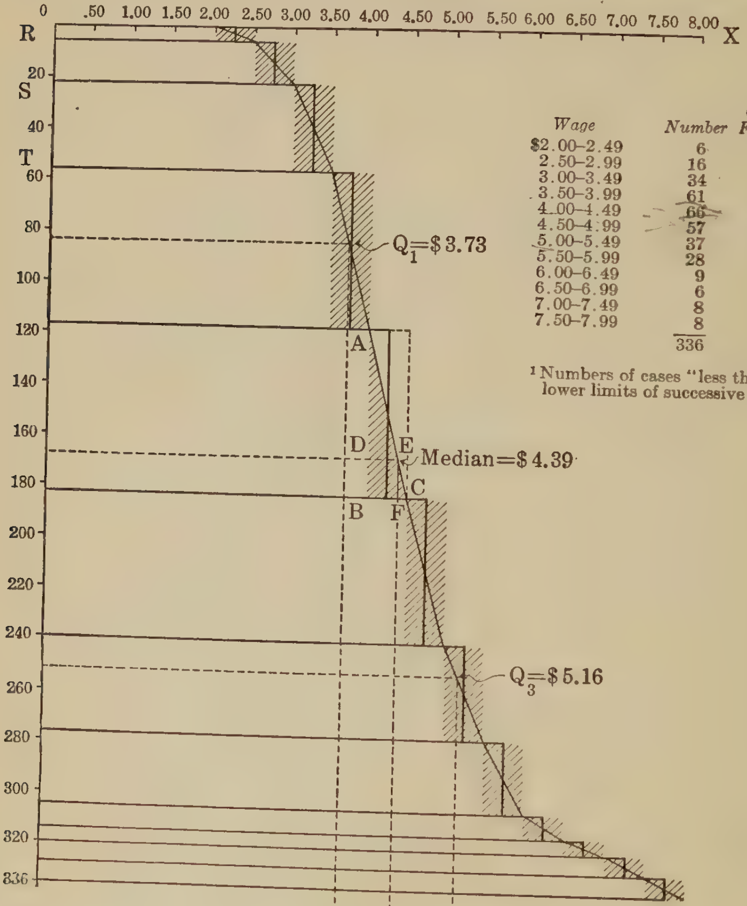
Ungrouped data. When the individual values are arranged in a simple ungrouped series in order of magnitude, the determination of the median is merely a matter of counting the items from the lowest or the highest until the single value is found which has an equal number of values above and below it, as already explained. When the number of items is even it is necessary to regard the median as located between the two middle values of the array. Unless these values are identical an average of the two mid-values must be obtained. *Furthermore, the student should always think of the median as a value rather than as a case. The location of the case is merely a means to obtain the value.*

The proportion method for grouped data. The data are very often found grouped in the form of a frequency distribution. In this situation it is not sufficient merely to locate the class in which the median falls, because this procedure determines no specific value but only a range of value between the limits of the class, 5 per cent, 10 pounds, or 50 cents. How then shall we determine with exactness the value of the median within the class-interval?

For purposes of locating the median within the class, the same assumption is made as in the computation of the mean. *The values in any class are assumed to be uniformly distributed throughout the range of the interval,* which amounts to the same thing as assuming concentration at the mid-value of the interval.

The location of the median exemplified in detail. The graphic representation of the frequency distribution on page 63 is repeated here for convenience. The student is asked to review the text on pages 64–65. We note from Figure 5 that the number of items is even (336), and our problem is to locate a value on the horizontal scale such that there will be as many value items above as below. Half the values would be $\frac{336}{2} = 168$ items.

Counting down on the vertical scale through the first four groups there will be cumulated 117 items. This brings us to the wage $4.00 on the horizontal scale. If we count in addition all of the fifth group, which has 66 items in it, there will be 183 items up to $4.50 on the horizontal scale.



Wage	Number	Cumulative Frequencies.[1]
$2.00–2.49	6	0
2.50–2.99	16	6
3.00–3.49	34	22
3.50–3.99	61	56
4.00–4.49	66	117
4.50–4.99	57	183
5.00–5.49	37	240
5.50–5.99	28	277
6.00–6.49	9	305
6.50–6.99	6	314
7.00–7.49	8	320
7.50–7.99	8	328
	336	336

[1] Numbers of cases "less than" the lower limits of successive classes.

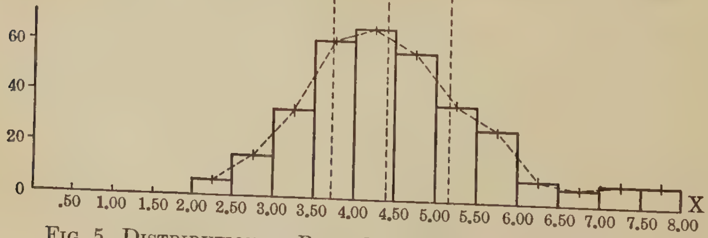


FIG. 5. DISTRIBUTION OF PIECE-RATE EARNINGS IN FIFTY-CENT GROUPS

FIG. 6. FREQUENCY HISTOGRAM AND POLYGON

But we wish a value such as will show only 168 items below and the same number above. Therefore, the median value must lie somewhere in the fifth group, between $4.00 and $4.50. This group, by assumption, is distributed evenly over the 50 cent interval. The lowest value in this group is at $4.00 and the highest just under $4.50, as shown by the diagonal AC and the distance BC in Figure 5. How many of the 66 items do we need, added to the 117 of the first four groups, to complete the 168 items which must lie below the median value of the entire array? There are needed 168 − 117 = 51 items to complete the 168, as shown by the distance AD measured on the vertical scale within the fifth group. The values in this group are increasing in amount as we count from A toward D, as shown along the diagonal AC. Beginning with $4.00 at A, the lower limit of the group, how much greater will the value be when we have counted 51 items along the vertical scale toward B? AB represents 66 items, BC represents 50 cents, and AD represents 51 items. What will be the value range represented by DE? The triangles ABC and ADE are similar. Therefore, the proportion $AB : BC :: AD : DE$ will be true. Substituting their specific values, we have, 66 : 50 cents : : 51 : DE, or 66 DE = 2550, and DE = 39 cents. Adding 39 cents to $4.00, the lower limit of the group, we have $4.39, the median value.

Or, reasoning in terms of proportions, on the assumption of even distribution, it follows that if 66 items are distributed over the entire range of 50 cents (BC), in the fifth group, 51 items, which must be added in order to have 168 items below the median value, will range over $\frac{51}{66}$ of 50 cents, or 39 cents (DE), which amount must be added to the lowest wage item of the group, $4.00, to arrive at the median value at E.

Exactly the same value for the median will be secured if we begin at the bottom of Figure 5, the highest wage values, and count upward on the vertical scale. As we proceed the wage values grow progressively smaller in each group. At the upper limit of the group of 66 items (C), where the value is $4.50, we have counted 153 items. It requires 15 more items in the group of 66 to fulfill the condition that 168 items must be greater than the median value. In the previous calculation we counted into this same group 51 items from the lower limit to arrive at the median. Now we count 15 items from the upper limit to arrive at exactly the same point. $AD = 51$, DB or $EF = 15$, and $AD + EF = 66$, the entire group. The wage values are now decreasing along the diagonal C to E, where the median is located. *The value to be subtracted* from $4.50 may be secured, as before, by the proportion, $AB : BC :: EF : FC$, or 66 : 50 cents : : 15 : FC, which gives 11 cents for the value of FC; or if 66 items range over 50 cents (BC), 15 items cover $\frac{15}{66}$ of 50 cents, or 11 cents (FC). Then the median value is \$4.50 − .11 = \$4.39.

The quartiles and their determination. The first quartile, designated Q_1, is that value which is exceeded by three fourths of the values and of which one fourth fall short. The third quartile, designated Q_3, is that value which is exceeded by one fourth of the values and of which three fourths fall short. In this problem, therefore, we divide the entire array of measurements into four equal groups, 336 ÷ 4 = 84 in each. To determine the first quartile wage we seek a value below which 84 items are located. Counting the first three groups from the lowest value of the array, 6 + 16 + 34, we have 56 items. The first quartile falls in the next group of 61 items which range over a 50-cent interval, from the lower limit \$3.50 to the upper limit \$4.00, as is evident in Figure 5. The difference between 84 and 56, or 28, gives the number of items in the group of 61, needed to fulfill the conditions of the definition of the first quartile. If the entire group of 61 items, in which the first quartile falls, ranges over a 50-cent interval, 28 items in that group will range over $\frac{28}{61}$ of 50 cents, or 23 cents, which must be added to the lowest value in the group, \$3.50, making the first quartile wage \$3.73. This is the wage below which are found 84, or one fourth of the wage items, and above which are located 252, or three fourths of the wage items.

The third quartile may be located from either limit of the array. Starting at the lower limit, we seek a wage value below which three fourths, or 252 items, are located. Counting through the first six groups, 6 + 16 + 34 + 61 + 66 + 57, we have 240 items. The third quartile falls in the next group of 37 items, which are distributed over a 50-cent interval, from the lower limit \$5.00 to the upper limit \$5.50. The difference between 252 and 240, or 12, indicates the number of items in the group of 37 needed to conform to the definition of the third quartile. If the entire group of 37 items, in which the third quartile falls, ranges over a 50-cent interval, 12 items of that group will range over $\frac{12}{37}$ of 50 cents, or 16 cents, which must be added to the lowest value in the group \$5.00, making the third quartile wage \$5.16. This is the wage below which are located 252, or three fourths of the wage items, and above which are located 84, or one fourth of the wage items.

The third quartile may be located also from the highest value in the array. We seek a value, exceeded by 84 items or one fourth of the entire array. We count $8 + 8 + 6 + 9 + 28 = 59$; $84 - 59 = 25$ items in the group of 37; $\frac{25}{37}$ of 50 cents = 34 cents, the range of value for 25 items. $5.50 − .34 = $5.16, the third quartile wage. This procedure is exactly the same as for the first quartile except that we count from a higher value toward a lower value and, therefore, subtract the 34 cents from the upper limit of the group in which the quartile is located, $5.50.

The same method has been used to locate the quartiles which was employed for the median. While the median divides the entire array of items into two equal parts, the quartiles divide each of these parts, in turn, into two equal parts, making four equal parts for the entire distribution of items. *Therefore, half the items in any distribution are located between the two quartiles.*

The significance of the position of the quartiles. *The median and quartiles give us three values for the more definite description of the structure of the frequency distribution, instead of a single value, as in the case of the mean.* In the wage problem one half of the items, or 168, are located within the range $3.73 to $5.16, a difference of $1.43. If the quartile values fall nearer to each other and to the median it signifies a greater concentration of the items about the central value, greater similarity or homogeneity of values. If the quartiles fall further apart it indicates the opposite tendency. Therefore, in describing the frequency distribution, the *position of the quartiles becomes a measure of similarity or homogeneity of values.* It should be pointed out that this measure of the degree of concentration of the data utilizes only the central half of the distribution and disregards the other half.

This method is useful in testing whether union organization tends to level out wages, that is, to make the wages in a given trade more nearly alike. If this be true, the quartile values in a wage distribution of union labor should more closely approach the average wage, showing a less tendency to deviate from this central wage value, than in the case of a wage distribution of unorganized labor.

The deciles and percentiles. The array of items may be divided into ten equal parts by a procedure exactly similar to that employed for the median and quartiles. For instance, a value may be located so that nine tenths of the items are greater and one tenth less in value. This is designated the first decile value. Or, a value may be located so that only one tenth of the items is greater and nine tenths less. This is designated the ninth decile value. In this manner the structure of the frequency distribution may be still more definitely described by points of reference in addition to the median and quartiles.

This procedure may be extended further by dividing the array into one hundred equal parts and locating percentile values by the same method as that used for determining quartiles and deciles. *The object of all these devices is to increase the definiteness and accuracy of our knowledge concerning the frequency distribution. It is not always safe to depend solely upon a single summary value, such as the mean or median, for the purpose of reducing data to simpler terms.*

The graphic method of locating median and quartiles. Figures 5 and 6 present graphically the facts of the problem used for illustration in the preceding discussion. The exact location within the class-interval of the median and quartiles was determined by the proportion method. These values may be located also directly from the diagram. On the vertical scale of Figure 5 each wage item is represented by an equal space, the size of the item being indicated at the same time on the horizontal scale. To locate the median graphically it is only necessary to count through 50 per cent or 168 of the spaces on the vertical scale and to draw a line horizontal to the base until it meets the diagonal line *AC* at *E*. Dropping a perpendicular from this point at *E* until it cuts the base line on the horizontal scale locates the median wage at about the same value as that more exactly determined by the proportion method at $4.39. This perpendicular also divides the area of the frequency histogram or polygon into two equal parts, as shown in Figure 6, and this is in accordance with the definition of the median.

To locate the first quartile graphically we need only count downward on the vertical scale through 25 per cent of the spaces, 84, and draw a horizontal line to meet the diagonal in the fourth group. The perpendicular dropped from this point intersects the horizontal scale at about $3.73, the first quartile wage, and divides the area of Figure 6 below the median into two equal parts, in accordance with the definition of the first quartile. Likewise, to locate the third quartile, count through 75 per cent, or 252, of the spaces and proceed as before. The perpendicular dropped from the point of intersection with the diagonal in the seventh group meets the horizontal scale at about $5.16, the third quartile wage, and divides the area of Figure 6 above the median into two equal parts. It will be noted that the median and quartiles, located in the manner described, *divide the area of Figure 6 into four equal parts.* The deciles and percentiles may be located graphically in the same manner and these divide the *frequency area* into ten and one hundred equal parts.

Warning should be given at this point *that it is not the base line distance of Figure 6 which is divided into four equal parts by the median and quartiles. It is the area of the frequency polygon,* which represents the entire number of items in the array. Frequency polygons with different shapes have different relative distances between the two quartile values, between either quartile value and the median, and between either quartile and the lowest or highest value of the entire distribution. This will be explained more fully in the discussion of variation and its measurement in Chapter IX.

It is suggested that the student return to the table on income distribution presented at the opening of this chapter and locate the median and quartiles by the proportion method.

USE OF THE CUMULATIVE FREQUENCY DISTRIBUTION

In Chapter V this method of presenting grouped data has been shown graphically in Figure 5. In the present chapter the same diagram has been repeated in order to explain the location of the median and quartiles by the graphic method (page 110). The purpose of the cumulative form of presentation in a table is to give the reader information at a glance as to the number and proportion of the total values in the array which are located above or below a given value in the distribution.

Table 20 presents the cumulative frequencies in two forms: (A) the

TABLE 20. CUMULATIVE FREQUENCIES AND PERCENTAGES

WAGE (class limits)	*f*	(A) "LESS THAN" (cumulative)		(B) "AT OR MORE THAN" (cumulative)	
		Number	Per cent	Number	Per cent
(1)	(2)	(3)	(4)	(5)	(6)
$2.00–2.49	6	0	0	336	100.0
2.50–	16	6	1.8	330	98.2
3.00–	34	22	6.5	314	93.5
3.50–	61	56	16.7	280	83.3
4.00–	66	117	34.8	219	65.2
4.50–	57	183	54.5	153	45.5
5.00–	37	240	71.4	96	28.6
5.50–	28	277	82.4	59	17.6
6.00–	9	305	90.8	31	9.2
6.50–	6	314	93.5	22	6.5
7.00–	8	320	95.2	16	4.8
7.50–7.99	8	328	97.6	8	2.4
		336	100.0		
	336				

total number of cases *below or "less than"* each successive value stated as the lower limits in column (1); (B) the total number of cases *"at or more than"* the value stated as the lower limits in column (1). The data are the same as in Figures 5 and 6. Not only are the cumulative frequencies given in the table, but also the cumulative percentages of the total items.

The percentages in columns (4) and (6) are not essential, but they do enable the reader to know at a glance what *proportion* of the total workers are earning *"less than"* or *"at or above"* $2.00, $2.50, and so on. *Columns (4) and (6) when added for any horizontal line always equal one hundred per cent.* For example, column (3) shows that 183 workers receive less than $4.50 which is 54.5 per cent of the total, as given in column (4). Likewise column (5) shows 153 workers receiving $4.50 or above, which is 45.5 per cent of the total, as recorded in column (6). This accounts for the total workers, 336, by describing them from the point of view of whether they earn less or more than $4.50. In the same manner it is possible to use the lower limit value of any class-interval and know at once the number and proportion below, or at or above, this limit.

The percentages also indicate the class-intervals in which the median and quartiles are located, although it is not possible to state their exact value. For example, 34.8 per cent receive less than $4.00 and 54.5 per cent receive less than $4.50. It follows that the median must be located at some value between $4.00 and $4.50.

In Figure 11 both series of cumulative frequencies from Table 20 are represented graphically on the same scales. The method of plotting is more direct than that used in Figure 5, and the diagrams serve equally well for the graphic location of the median and quartiles.

The vertical and horizontal scales are the same as in Figure 5, and diagram (A) is exactly similar to the part of Figure 5 which is included within the continuous diagonal lines drawn through the successive cross-hatched areas from $2.00 to $8.00 on the horizontal scale and from zero to 336 on the vertical scale. It will be noted that the zero line of Figure 11 is located at the top of the diagram to correspond with Figure 5.

In plotting the successive cumulative frequencies to form diagram (A), from column (3) of Table 20, *care must be taken to locate them at the limit of each class-interval,* rather than at the middle as in the case of the frequency polygon, *in order to include all cases less than the given value.* For example, at less than $2.00 there are no items. Therefore, the first point is located at $2.00 on the horizontal scale and on the line drawn through zero on the vertical scale. Directly opposite $2.50, a distance equal to six items on the vertical scale is laid off and the point is located by a dot (.). Opposite $3.00 a distance is measured representing 22 items on the ver-

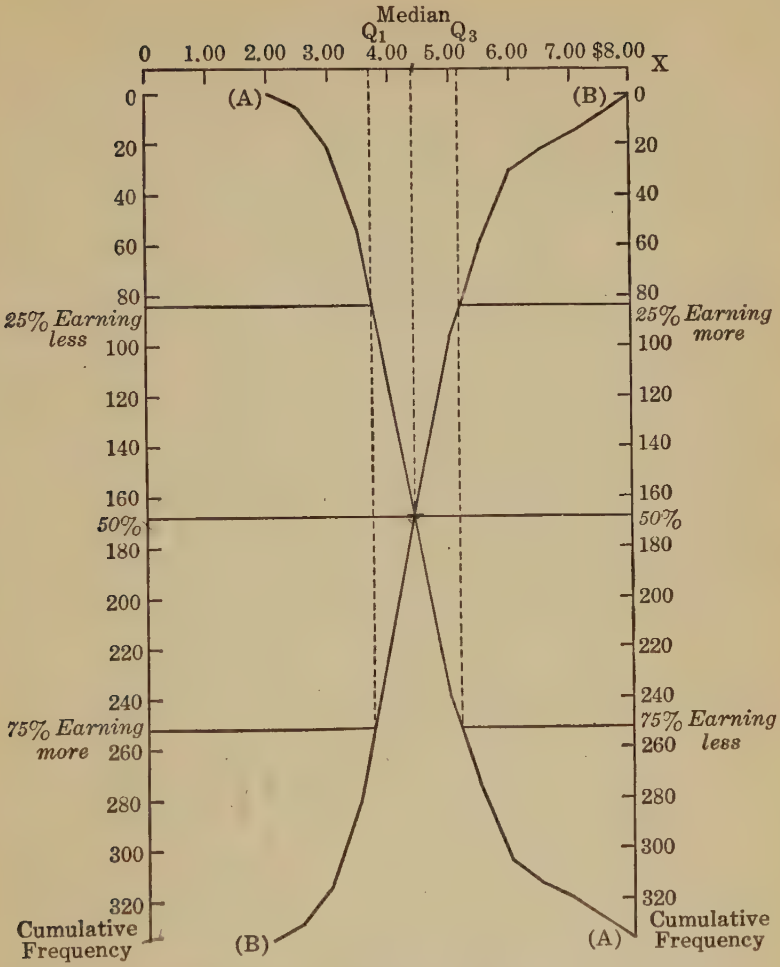


FIG. 11. REPRESENTATION OF CUMULATIVE FREQUENCY DISTRIBUTIONS
Graphic location of Median and Quartiles. (Data from Table 20.)

tical scale, and so on for each successive cumulative frequency. Finally, a distance is measured on the vertical scale equal to the total cases in the distribution, 336, and the point is located opposite $8.00. The dots are then connected by straight lines, forming a cumulative frequency diagram (A). If the class-intervals were indefinitely narrowed and if the number of cases were correspondingly increased the diagram would become *a smoothed cumulative frequency curve.*

Likewise, diagram (B) is plotted from column (5) of Table 20. In this case the first dot is located directly opposite $2.00, the lower limit of the first class-interval, in order to include all cases at $2.00 or above. The vertical distance represents 336 items, the total cases in the distribution. The next point is located opposite $2.50, a distance on the vertical scale equal to 330 items, all but six in the entire series. This procedure is continued for each cumulative frequency in colunm (5), in the reverse order to that in (A). A final point is located at $8.00, since there are no items above this value. Again the dots are connected by straight lines, to form diagram (B).

It will be observed that the vertical scale is repeated on the right of the diagrams for convenience in locating the median and quartiles. To determine the median from either (A) or (B) it is only necessary to locate the point on the vertical scale having an equal proportion of the items above and below it, and to draw a horizontal line through this point until it meets the curve. If (A) and (B) are correctly plotted they must intersect at a point directly opposite the median value on the horizontal scale, since it must be possible to determine the median from either. The dotted vertical line drawn through the point of intersection meets the horizontal scale at the median value, about $4.39.

The quartiles also may be located from either (A) or (B). For this purpose the horizontal lines are drawn through points on the vertical scale representing 25 per cent and 75 per cent of the items until they intersect (A) or (B). The corresponding vertical dotted lines are drawn through the points of intersection, to meet the horizontal scale at about $3.73 and $5.16. The "25 per cent earning less" marked on the left-hand scale means that 25 per cent of the cases are at lower values and refers to (A); while the "75 per cent earning more" means that 75 per cent are at higher values and refers to (B). On the right-hand scale the " 25 per cent earning more " means that 25 per cent are at higher values and refers to (B); while "75 per cent earning less" means that 75 per cent are at lower values and refers to (A).

The *graphic method* of locating the median and quartiles is usually not so accurate as the *proportion method*. In practice only one of the diagrams, (A) or (B), would be drawn for purposes of locating the median and quartiles.

Again the reader is asked to remember that the fundamental purpose in presenting these devices is to describe the structure of the frequency distribution as definitely as possible for purposes of analysis and comparison.

The use of cumulative frequency curves for interpolation. When the cumulative frequency diagrams (A) and (B) have been smoothed to form *cumulative frequency curves*, the class limits, with which we began in the original distribution, are no longer important. It becomes possible from such a curve to find the number of wage-earners in any desired interval on the wage scale, even though this interval is not the same as any of those in the original frequency distribution. For example, if we desire to find out the number earning \$4.25 to \$4.75, we locate \$4.25 on the horizontal scale and read on the vertical scale from curve (A) immediately opposite this value the number receiving less than \$4.25. Similarly we locate \$4.75 and read from the curve the number earning less than this amount. The difference between these two readings on the vertical scale is the number of wage-earners included in the interval \$4.25 to \$4.75. Securing values between those given by the original tabulation is called interpolation. We are already familiar with this principle, illustrated in the location of the median and quartiles within a class-interval.

THE EFFECT OF DIFFERENT SIZED CLASS-INTERVALS UPON MEDIAN AND QUARTILES

Using the frequency table of the weights of college Freshmen, page 100, in ten-pound groups, let us determine the median and quartile weights. One half the measurements would be 500, one quarter, 250, and three quarters, 750. Counting through from lower to higher values:

(A) $13 + 28 + 146 = 187$.

$250 - 187 = 63$, needed in interval 120–130 pounds to make 250 items.

$\frac{63}{245}$ of 10 pounds = 2.6 pounds, to be added to 120 pounds.

$120 + 2.6$ pounds = 122.6 pounds = *first quartile value.*

(B) $13 + 28 + 146 + 245 = 432$.

$500 - 432 = 68$, needed in interval 130–140 pounds to make 500 items.

$\frac{68}{242}$ of 10 pounds = 2.8 pounds, to be added to 130 pounds.

$130 + 2.8$ pounds = 132.8 pounds = *median.*

(C) $13 + 28 + 146 + 245 + 242 = 674$.

$750 - 674 = 76$, needed in interval 140–150 to make 750 items.

$\frac{76}{160}$ of 10 pounds = 4.8 pounds, to be added to 140 pounds.

$140 + 4.8$ pounds = 144.8 pounds = *third quartile value.*

It is suggested that for practice the student calculate the median and quartiles from the five-pound grouping on page 98. Differences will be found in the values obtained from the two types of grouping. The cause of these differences is found in the assumption of even distribution over the given class-interval, whether it be five or ten pounds. In a frequency table the items tend to mass toward the central value and, therefore, for any single interval of the distribution, *the items are more numerous in that half of the interval which is located nearer the central value.* The table showing five-pound groups makes this clear. Double the five-pound intervals and examine the frequencies, for example, from 120 to 130 pounds, the interval within which the lower quartile is located. There are 111 items in the lower half, 120 to 125 pounds, and 134 items in the upper half, 125 to 130 pounds. After we pass the central value of the distribution, the opposite tendency may be observed. In the interval 140 to 150 pounds, in which the upper quartile is located, there are 85 items in the lower half of the interval and only 75 items in the upper five-pound group.

Therefore, the assumption of even distribution over the ten-pound interval is not so exact as the same assumption over the five-pound interval. By using a five-pound interval, a somewhat truer picture of the facts is secured than by a grouping twice as large, because the assumption in the case of the narrower grouping is subject to less error. This accounts for the fact that the first quartile, for the five-pound grouping, is a higher value and the third quartile a lower value than for the wider interval. *The ten-pound grouping exaggerates the spread between the two quartiles by about a half-pound.*

Experiments such as these suggested demonstrate the importance of the size of the class-interval, discussed in Chapter V. In the calculation of the mean, it has been shown that the difference is insignificant. The same is true for the median. If, however, we use the quartiles to indicate the variation about the average, then the use of the five-pound grouping has the advantage of greater accuracy.

Comparison of the mean and median weight. In this instance the median is about two pounds less than the mean, which indicates the influence of the higher values upon the latter. The frequency table shows that the value about which the greatest density of items occurs is not in the middle of the entire range of values, but nearer the lower limit. The higher values, rapidly decreasing in number, tend to pull the mean upward in the scale, but do not affect the median nearly so much, *since it is merely the number of the items, not their size, which influences the position of the latter.* In fact, it is not necessary to know the exact values of extreme variants at either end of the distribution in order to calculate the median, provided always that we know whether they are above or below the central value, and how many such items there are. This must be true from our definition of the median as a *position average.*

When the quartiles are used with the median it is possible to describe the distribution much more definitely than by the use of a single summary figure, the mean. For example, the quartile weights are located about 22 pounds apart. Within this range of value one half, or 500, of the weight items are concentrated. The other 500 vary over a range from 90 to 122 pounds and from 145 to 210 pounds, a total of only a little less than 100 pounds. We are utilizing the central half of the distribution to inform ourselves about the extent of the concentration at or near the central value.

It is evident that there is some risk in describing an entire distribution by the characteristics of the middle half of it. Whether this method leads to serious error may be checked readily by reference to the detailed frequency table or the graphic representation.

Computation from an odd number of cases. A final illustration presents a frequency table in which the mid-values are stated rather than the limits of the class-intervals and in which the number of items is odd rather than even.

TABLE 21. COMPUTATION OF MEDIAN AND QUARTILE AGES

(Age to the nearest birthday)

AGE (years) m (1)	f (2)
5	4
6	9
7	50
8	86
9	54
10	24
11	13
12	10
13	5
Total....	255

One half of the items is $127\frac{1}{2}$, one fourth $63\frac{3}{4}$, three fourths $191\frac{1}{4}$, counting from the lowest age. To find the median value, we add $4 + 9 + 50 = 63$; $127\frac{1}{2} - 63 = 64\frac{1}{2}$ needed to make $127\frac{1}{2}$ items; $\frac{64\frac{1}{2}}{86}$ of 1 year = .75 year. This .75 year must be added to the lower limit of the fourth group, in which the median is located. What are the limits of this group? Since the age is stated *to the nearest birthday*, rather than to the last birthday, the 86 children in this class are distributed a half-year above and a half-year below eight years. In other words, children are classed at eight years who have just reached 7.5 years and who have not yet reached 8.5 years. Therefore, .75 year is added to the lower limit of the interval, 7.5 years, which makes 8.25 years, the median age. *The student is cautioned not to add .75 year to the mid-value of the class.*

The quartiles are located in a similar manner.

(A) $4 + 9 + 50 = 63$; $63\frac{3}{4} - 63 = \frac{3}{4}$ item, needed to make $63\frac{3}{4}$.

$\frac{\frac{3}{4}}{86}$ of 1 year = .01 year, to be added to 7.5 years, the lower limit of the interval.

7.5 years + .01 year = 7.51 years = Q_1.

(B) $4 + 9 + 50 + 86 = 149$; $191\frac{1}{4} - 149 = 42\frac{1}{4}$ items, needed to make $191\frac{1}{4}$.

$\frac{42\frac{1}{4}}{54}$ of 1 year = .78 year to be added to 8.5 years, the lower limit of the interval.

8.5 years + .78 year = 9.28 years = Q_3.

When dividing the number of items by two and by four, to locate the median and quartile positions in the entire array, fractions sometimes result, as in the problem. The student is advised to use these fractions of items in locating the median and quartile values within the class-interval, not because it is usually important from the point of view of the accuracy of the results, but because, otherwise, it will not be possible to obtain exactly the same values for median and quartiles if in checking our work we count from the opposite end of the array.

As illustrated also by this problem, it is worth noting that sometimes the manner of making the measurements or collecting the data establishes the class-interval and its mid-value. For example, the requirement of age to the nearest birthday determines at once the class-interval of one year and makes it necessary to record the mid-value of the class. A similar procedure is followed in measuring height to the nearest quarter-inch. In this case there is a variation of an eighth of an inch on either side of the value actually recorded. The class-interval is a quarter-inch and the lower limit of the class is one eighth below the recorded amount, and the upper limit is one eighth above the recorded value. All cases measured within that range are recorded at the given quarter-inch.

APPLICATIONS OF THE MEDIAN

Compared with the mean, the median has a more restricted application in practice. However, it seems desirable to promote a better understanding of this form of average among the consumers of statistics, because many distributions can be well characterized by it, and with less labor of calculation, while for other distributions its use is essential in the interest of accuracy. When it is used with the quartiles we have a description of the distribution about the central value which is essential in judging the significance of the average.

The median has been widely used recently in the presentation of wage statistics. This form of average minimizes the influence of extreme variations and often gives a truer picture than the mean, especially when the quartiles are introduced as additional values of reference. It has proved a useful average, also, in the construction of index numbers describing variations in wages and commodity prices, where it is desired to minimize the influence of extreme fluctuations.

The position of the quartiles, near or far from the median, becomes a valuable index of homogeneity or likeness of at least the central half of the values. By the interquartile range it is possible to describe the usual or normal variability in the weights of school children of the same age and height. This information is important in deciding how far a particular child may be permitted to fall below accepted standards based upon average weight, for a given age and height, before he is diagnosed as abnormally underweight. How much variation in weight is to be expected among healthy children is a matter of experiment and measurement.

Those who deal with anthropometric measurements find the median useful in quickly characterizing a series of data. The distribution is likely to be regular with a very definite massing of the items about a central value. Both the mean and the median are satisfactory representative values but the latter is more easily determined.

The median is used also in the field of population statistics.[1] When the ages of those living at a given time are classified in single year groups, or longer intervals of five or ten years, the median may readily be determined and may be employed to describe the age distribution. This value may be called the average or "central age" or the "probable age" of those living at any given time. The chances are even that the age of any individual chosen at random from this population group will be greater or less than the median age. This measure is also used to express the "probable lifetime" and is computed from the lengths of life in a mortality table, for a given generation. It expresses the age which half of the persons of a given generation, born at the same time, will survive. The "probable lifetime" differs from the "expectation of life," for the computation of which the mean is used. In like manner, by determining the median of the ages of persons marrying the "probable age of marriage" may be described.

[1] Cf. Zizek, *Statistical Averages*, translated by Warren M. Persons, pp. 215–17.

SUMMARY OF THE MEDIAN AVERAGE

In the light of the criteria according to which an average may be judged and which are enumerated at the close of the preceding chapter, we may make the following observations concerning the median.

1. In contrast with the mean, the median is not so satisfactory from the point of view of arithmetical and algebraic treatment. For example, the total of the values cannot be computed from the median by multiplying it by the total frequencies, as can be done in the case of the mean.

2. It is a position average and, therefore, the data must first be arrayed according to size before it can be determined.

3. This measure is relatively independent of the influence of extreme variations, since the size of these items does not affect it. Variations in the individual measurements do not influence it, provided they do not affect the central items of the distribution, and provided the number of items above and below the central value remains unchanged.

4. Undistributed groups at the extremes of the distribution offer no difficulty in locating the median so long as the number of items is known, for example, in a wage distribution, a group earning less than $10 or a group earning $25 or more. In the case of the mean these undistributed groups introduce an element of uncertainty in the computation.

5. The median may not prove to be a representative value. It may fall where there are few items in the series. In this case the median should be abandoned, as well as the mean. Some other device, as the frequency polygon, must be used to describe the distribution.

6. It is more easily determined than the mean but it is not so widely understood.

7. When used with the quartiles the median furnishes a more definite description of the structure of the distribution than the mean alone.

The median is a useful descriptive value, where the data are grouped with a fair degree of regularity around the central value, and where some few extreme variants exercise undue influence on the mean; where both mean and median are typical, but the latter has the advantage of quicker determination; where the median and quartiles offer a more complete characterization of the structure of the distribution around the average and enable the student to understand more clearly the significance of the average; or where a part of the data are not subject to exact individual determination.

THE GEOMETRIC MEAN

The present chapter has been devoted to a discussion of the median, its determination and its uses. It has been strongly emphasized that this form of average minimizes the influence of extreme variants. Another form of average, the geometric mean, not so widely used or understood as either the mean or the median, has this same characteristic. The geometric mean of N items is the Nth root of the product of the values. For example, if we combine two items, 160 and 250, by this method we have $\sqrt{160 \times 250} = 200$. The arithmetic mean of these values is 205.

Like the arithmetic mean this form of average is influenced by the size of all the items in the series and may not exactly coincide with any of them. However, it is less affected by extreme values than the mean. It is never greater than the mean, usually only slightly less. At present it is not widely employed in the social sciences or in business, but it has been applied in estimating population growth and in constructing index numbers. Jevons,[1] who first used this method in computing the mean index of commodity prices, has been followed in this field by recent writers and research workers. The utility of the geometric mean in the computation of index numbers will be discussed in Chapter X, when that topic is considered.

Computation of the geometric mean. It is sometimes called the *logarithmic average* because logarithms are used in its computation. *It is the natural number which corresponds to the arithmetic mean of the logarithms of the individual values in the series to be averaged,*

$$\text{log of Geometric Mean} = \frac{\Sigma \log X}{N}.$$

This formula describes the procedure to be followed when any number of values are combined in a geometric mean. In a preceding paragraph only two values were averaged. In this case the two values were multiplied and the square root was extracted. But to average a larger number of items, requires the extraction of the *Nth* root of the product of all the values. The use of logarithms makes this procedure easy. The logarithm of the product of all the items is obtained by adding together the logarithms of the separate items. The logarithm of the *Nth* root is obtained by dividing this sum by N. *The result is the logarithm of the desired geometric mean.* It only remains to look up in a table of natural and logarithmic numbers *the natural number which corresponds to the logarithm of the geometric mean. The result is the geometric mean.*

[1] Jevons, in the third edition (1888) of his *Theory of Political Economy*, p. xxix, referring to Cournot's treatise, says, "The second chapter contains an important anticipation of discussions concerning the proper method of treating prices, including an anticipation of my logarithmic method of ascertaining variations in the value of gold."

The use of the geometric mean to measure the rate of population growth. A census of population is taken every ten years by the Federal Census Bureau, the most recent count being on January 1, 1920. We shall assume the population of a certain city to have been exactly 100,000 in 1910. In 1920 the population of this city had increased to 150,000, a growth of fifty per cent in ten years.[1]

We wish to find the constant annual rate of growth in order to estimate the population of this city at inter-census and post-census years. For example, the health authorities wish to know the population of 1911, when no count was made, in order to calculate death-rates and birth-rates for that year; or the real estate interests wish to project the growth beyond the last census of 1920, in order to estimate the probable population of 1924. If the yearly rate of growth is known the population of the inter-census years can be computed on the principle of compound interest, year by year, starting with 100,000 in 1910. Or, if the same rate is assumed to continue beyond 1920, similar calculations may be made for post-census periods starting with 150,000 in 1920.

What is the average annual rate of growth if the increase has been fifty per cent in ten years? Is it permissible to divide fifty per cent by ten and obtain an annual rate of five per cent? The student should use five per cent as the rate, beginning with 100,000 and compound the increase annually at this rate for ten years, exactly as in compound interest. The result is not 150,000 in 1920, but 162,891.

In order to determine a constant annual rate of growth during a decade, the principle of geometric progression is employed. The procedure may be generalized as follows:

To find the rate of growth.

Let P_0 = population in 1910 = 100,000
Let P_1 = " " 1920 = 150,000
Let r = rate of growth
Then, population of 1911 $= P_0 + P_0 r = P_0\,(1 + r)$,
population of 1912 $= P_0(1 + r)\,(1 + r) = P_0(1 + r)^2$,
and so on for each year of the decade until finally,
population of 1920 $= P_0(1 + r)^{10} = P_1 = 150{,}000.$ (1)
Using logarithms for the computation of r, we have
$\log P_0 + 10 \log(1 + r) = \log P_1$, which is equivalent to equation (1) above, and transposing and dividing by 10,

$$\log\,(1 + r) = \frac{\log P_1 - \log P_0}{10}$$

[1] The date of the 1910 Census was April 15, 1910. The elapsed period between counts was actually 9.71 years, but for the sake of greater simplicity we assume ten years.

Substituting known values for P_1 and P_0,

$$\log (1 + r) = \frac{\log 150{,}000 - \log 100{,}000}{10}$$

$$= \frac{5.17609 - 5.00000}{10}$$

$$= .017609$$

Therefore, $(1 + r) = 1.04138$ [1]

and $r = .04138$ or 4.138 per cent.

$$\frac{5.17609 - 5.00...}{}$$

0.17609

$\therefore (1+r) = 1.04138$

Applying this rate of 4.138 per cent to the population of 100,000 and compounding it annually for ten years produces 150,000, the actual counted population of 1920. We are now in a position to estimate the population at any year between the two censuses, or to continue the same percentage growth beyond 1920, *on the assumption of a uniform rate.* This is only one method of estimating growth, but it is a widely used method.

READINGS

Elderton, W. P., and Ethel M., *Primer of Statistics*, chaps. 1 and 2.
Rugg, H. O., *Statistical Methods Applied to Education*, chap. 5.
Mills, F. C., *Statistical Methods Applied to Economics and Business*, chap. 4.
King, W. I., *Elements of Statistical Method*, chap. 12.
Yule, G. U., *An Introduction to the Theory of Statistics*, 6th ed., chap. 7.
Bowley, A. L., *Elements of Statistics*, 4th ed., part I, chap. 5.
Jerome, Harry, *Statistical Method*, chap. 7.
Secrist, Horace, *An Introduction to Statistical Methods*, chap. 8.
—— ——, *Readings and Problems in Statistical Methods*, chap. 7.
Zizek, Franz, *Statistical Averages*. Translated by Warren M. Persons, part II, chap. 3 ("Geometric Mean"), and chap. 4 ("The Median").
Jones, D. C., *A First Course in Statistics*, chap. 4.

REFERENCES

Kelley, Truman L., *Statistical Method*, chap. 3.
Rietz, H. L. (Editor), *Handbook of Mathematical Statistics*, chap. 2, prepared by Professor Rietz.
Whipple, G. C., *Vital Statistics*, 2d ed., chaps. 5 and 6. (Application of geometric principle in estimating mean population at inter-censal and post-censal periods.)
Newsholme, Arthur, *The Elements of Vital Statistics*, new ed., 1924, chap. 3. (Application of the geometric average in estimating population.)

[Details as to publisher and date of publication of the readings and references given at the close of each chapter are found in the alphabetical list in Appendix A.]

[1] The logarithm of $1 + r = .017609$. It is necessary to look up, in a table of natural and logarithmic numbers, the natural number (1.04138) which corresponds to the logarithm .017609.

CHAPTER VIII

METHODS OF SUMMARIZATION AND DESCRIPTION — THE MODE

The mode is the value which occurs most frequently and around which other items cluster most densely. In ungrouped data it is easily located by inspection. This value sometimes describes a distribution in a more representative manner than other forms of average.

In the frequency histogram on page 63, Figure 6, the *modal class* at $4.00 to $4.50 is indicated by the maximum height of the rectangle. More wage earners are grouped in this class than within any other 50-cent interval in the distribution. In other words, if any one of the 336 items were selected at random there would be greater probability of its belonging to this class than to any other in the entire distribution. A more definite value for the mode is often determined within the modal class by taking the mid-value, in this case $4.25. This is designated the *crude mode* as distinguished from the *true mode*, which will be explained later. In Figure 6 the crude mode is indicated by the highest point of the frequency polygon. In any case, the mode is a real value which is identical with actual measurements. Any refinement which neglects this consideration destroys the utility of this measure of central tendency.

In contrast with the mean, the *approximate mode* is determined without computation after the data have been grouped in order of size. It readily lends itself to graphic representation and, therefore, proves especially useful both in emphasizing certain characteristic parts of a distribution and in the study of specific aspects of a problem. The size of the mode does not depend upon the size of all of the items. *It is a position average.* Extreme variants and fluctuating values in other parts of the series than at the point of concentration have no effect upon it. This form of average used to characterize a distribution is in many cases the most logical of all, but it has definite limitations and should be employed with caution.

The mode is the form of average best understood by the layman, although he may not call it by this name. It is variously known as the predominant, the usual, the typical, the normal value. In any case it is understood to be the value most frequently found in an array of values.

Possibility of more than one mode. If a frequency distribution consists of a sufficient number of observations to be representative of the situation described, there is usually some degree of concentration of items at one or more points. *In the same distribution more than one mode may appear.* If these marked concentrations are really typical, and not merely accidental, a fact which may be experimentally tested by more observations of similar data, it generally signifies that the series is made up of dissimilar elements, each of which possesses a different central value around which the items tend to cluster. The diagram on page 104, for Factory II, represents this situation, where wages of both men and women are classified in the same distribution. If the investigator has access to the original data, the wage items can be re-classified so as to separate the dissimilar factors into more than one frequency distribution. When this has been done each distribution may be found to be regularly grouped about a single mode.

The heights of recruits for military service are likely to show more than one mode when different nationalities with different typical heights, as Italians and native Americans, are included in the same series. In certain occupations both the young and the old are likely to be selected for the same kind of work. In consequence the age grouping of the workers will have two modes. When the wages of all male workers in an industry are classified in a single series, without careful distinction as to the specific kind of work performed, more than one mode may appear because of the different typical earnings in different kinds of work.

If it is not possible to analyze a multi-modal distribution, *the investigator at least can avoid trying to summarize all the values by the computation of the median or the mean.* Graphic presentation of such a distribution is desirable.

THE DETERMINATION OF THE MODE

In an array of ungrouped data the mode is the value most frequently repeated and is easily determined by inspection. In grouped data the location of the class having the largest number of items is, likewise, a simple matter if we accept a specific size of interval in the grouping. The frequency table and histogram reveal the *modal class* on inspection. However, other sizes of interval with different class limits are possible. What happens when we re-classify the original data according to a different sized interval, or shift the limits up or down the scale? Reference to Figure 8 A and B, page 69, in which income data are grouped in two different intervals of $100 and $200, furnishes an answer. The *crude mode* falls at $950 in Figure 8 A and at $900 in Figure 8 B.

The weights of college freshmen, Table 3, page 57, are grouped in three different intervals, five pounds, ten pounds, and fifteen pounds.

The frequency polygons portraying these distributions are presented in Figure 7 A, B, and C, page 67. The limits of the *modal class* are 125 to 130 pounds for the five-pound grouping; 120 to 130 pounds for the ten-pound grouping; and 120 to 135 pounds for the fifteen-pound distribution. The *crude mode* is located at 127½ pounds, 125 pounds, and 127½ pounds in Figures A, B, and C respectively. It shifts to a lower position and back again as we enlarge the class-interval.

Indefinite location of the mode in grouped data. It is evident that the exact location of the mode, the massing point of the distribution, is not so simple to determine as at first it seemed. The *crude mode* is affected by the position of the class limits and the width of the interval chosen for classification of the data. The choice of interval is an experimental procedure as explained in Chapter V. The assumption of even distribution within any given interval is more or less arbitrary and not entirely true to the facts, as has been explained before. Therefore, varying the size or shifting the limits of the classes causes more or less change in the points of greatest density. Nevertheless, among the individual cases there is a value which appears most frequently, *a true mode.* It is our object to locate this value as nearly as possible. From the illustrations the reader should understand why the mode is sometimes called the "approximate inspection average," or an "unstable average."

The investigator may not have access to the original individual measurements. The data may come to his attention already grouped in a frequency table. In such a table the location of the *crude mode* by inspection may appear uncertain as the investigator examines the class in which the greatest number of items falls and the classes directly above and below. What experimentation may be employed to test the position of the mode and to locate it somewhat more exactly?

Table 22 presents four different class-intervals, five, ten, fifteen, and twenty pounds. The frequencies in the various columns are placed opposite the mid-values of the classes to which they relate. The lower limit of the first class in column (4) begins at 95 pounds, omitting the first five-pound group, and the data are grouped in ten-pound intervals. Two such shifts are made for the fifteen-pound grouping in columns (6) and (7), beginning the classes at 95 and 100 pounds, respectively, merely to show the effect on the mode when the lower limit of the distribution is shifted.[1] Since the mode is a position average, characterizing only the central part of the distribution, this omission of items on either margin of

[1] Only one shift of the lower limit of the distribution in twenty-pound groupings is made. As a rule, all possible shifts should be made for each new grouping. Two more could be made for the twenty-pound grouping.

TABLE 22. LOCATION OF THE MODE BY DIFFERENT GROUPINGS

WEIGHT (pounds) 5-pound interval		TEN-POUND GROUP		FIFTEEN-POUND GROUP			TWENTY-POUND GROUP	
	f	10-pound interval *f*	Shift one interval (beginning 95 pounds) *f*	15-pound interval *f*	Shift one interval (beginning 95 pounds) *f*	Shift two intervals (beginning 100 pounds) *f*	20-pound interval *f*	Shift one interval (beginning 95 pounds[a]) *f*
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
90 to 95	6	13	(Omit 6)	23	(Omit 6)	(Omit 6)	41	(Omit 6)
95 " 100	7		17		35	(Omit 7)		100
100 " 105	10	28				93		
105 " 110	18		83	164				
110 " 115	65	146			257		391	
115 " 120	81		192			326		451
120 " 125	111	245		370				
125 " 130	134		259		376			
130 " 135	125	242				327	402	
135 " 140	117		202	277				331
140 " 145	85	160			214			
145 " 150	75		129			164		
150 " 155	54	89		114			135	
155 " 160	35		60		81			94
160 " 165	25	46				59		
165 " 170	21		34	39				
170 " 175	13	18			23		27	
175 " 180	5		10			14		16
180 " 185	5	9		11				
185 " 190	4		6		7			
190 " 195	2	3				3	4	
195 " 200	1		1	2				2
200 " 205	0	1			(Omit 0)			
205 " 210	1		(Omit 1)		(Omit 1)	(Omit 1)		
Total	1000	1000	1000	1000	1000	1000	1000	1000

[a] As a rule the adoption of a larger class-interval should be accompanied by all possible shiftings of the limits of the classes. Two more shiftings of the lower limit of the first interval are possible with the twenty-pound grouping, that is, beginning at 100 pounds and at 105 pounds.

the entire distribution ought not to affect its position. But shifting the lower limit of the first class does affect the position of the modal class, and changes the position of the *crude mode* in the table. This follows from our assumption of even distribution over whatever class-interval is used. If the data were arranged in order of size but ungrouped, and items were omitted on the margins, it would not affect the position of the modal value.

The bold faced type used for the largest frequency in each column indicates the position of the *modal class* as it shifts about. This frequency occurs directly opposite the mid-value of the class to which it refers, as determined from column (1). Let us examine the range of value over which the mode shifts in columns (2) to (9).

Summary of Results from Table 22

Column Designation of Table 22	Modal Class
(2)	125 to 130 pounds
(3)	120 " 130 "
(4)	125 " 135 "
(5)	120 " 135 "
(6)	125 " 140 "
(7)	130 " 145 "
(8)	130 " 150 "
(9)	115 " 135 "

125–130 𝍸 / = 6 times
130–135 𝍸 / = 6 times

From first inspection of the five-pound grouping in Table 22 the two most likely intervals for the location of the mode appear to be 125 to 130 pounds and 130 to 135 pounds. Of the eight locations for the mode shown above how many are common to each of these two groups? Six of the intervals given in the summary of results from Table 22 include the range 125 to 130 pounds and an equal number, six, include the range 130 to 135 pounds. This situation suggests that the true mode probably falls at or near 130 pounds, between the two most probable class-intervals. Other evidence that this is so will be presented in the next section of this chapter.

Smoothing the distribution by a moving average of frequencies. The frequency polygon may be smoothed by taking the mean of two or more class-frequencies successively, dropping one class-frequency and adding another each time the mean is taken. In this manner the average is progressive through the distribution, keeping the interval the same size, and is called a *moving average*. This method is applied in Table 23.

In Table 23 the frequencies of two classes at a time are averaged; for example, $\frac{6+7}{2} = 6.5$. This mean frequency is entered in column (3) opposite 95 pounds, the class limit between the two intervals whose fre-

Table 23. The Moving Average in Locating the Mode — Smoothing the Distribution

Weight (pounds) (1)	*f* (2)	Two-Class Moving Average (3)
		3.0[a]
90 to 95	5	
		6.5
95 " 100	7	
		8.5
100 " 105	10	
		14.0
105 " 110	18	
		41.5
110 " 115	65	
		73.0
115 " 120	81	
		96.0
120 " 125	111	
		122.5
125 " 130	134	
		129.5
130 " 135	125	
		121.0
135 " 140	117	
		101.0
140 " 145	85	
		80.0
145 " 150	75	
		64.5
150 " 155	54	
		44.5
155 " 160	35	
		30.0
160 " 165	25	
		23.0
165 " 170	21	
		17.0
170 " 175	13	
		9.0
175 " 180	5	
		5.0
180 " 185	5	
		4.5
185 " 190	4	
		3.0
190 " 195	2	
		1.5
195 " 200	1	
		.5
200 " 205	0	
		.5
205 " 210	1	
		.5[a]
Total	1000	1000.0

[a] 3.0 and .5 at the margins of the distribution are secured by extending one interval where there are no cases and dividing 6 by 2 and 1 by 2. This keeps the total frequency 1000.

quencies are averaged. Then the first frequency is dropped and another added, $\frac{7+10}{2} = 8.5$. This is entered in column (3) opposite 100 pounds. The same method is used throughout the distribution. The fractional frequencies are preserved because we wish merely to smooth out any irregularities without changing the total frequencies. It is necessary to add a class, 85 to 90 pounds, at the lower end of the distribution, in which there are no cases, in order to obtain the first mean frequency, $\frac{0+6}{2} = 3$. Likewise, a class is added at the upper end of the distribution, 210 to 215 pounds, to obtain the last mean frequency, $\frac{0+1}{2} = .5$. Thus the total of the frequencies in column (3) is 1000.

This method of *moving average* locates the maximum frequency opposite 130 pounds, with a fairly symmetrical distribution above and below it. These facts are presented in Figure 12, page 136. By this smoothing process the mode is located at 130 pounds, which is about the same value as that obtained by the method of grouping and shifting in Table 22. Both methods are used as expedients in lieu of the more technical methods of curve-fitting.

The relative positions of mean, median, and mode. So far we have explained three forms of average. It will be useful to illustrate by diagrams the position of each in relation to the others in any given distribution. It is apparent that frequency distributions have different internal structures and various shapes when presented graphically. Some are almost perfectly symmetrical around the central value, others are moderately asymmetrical, and still others are decidedly asymmetrical. Any of these forms might be the *usual or normal form for the given type of data* provided a sufficient number of cases has been included to smooth out accidental irregularities. For example, the heights of army recruits are grouped with almost perfect bell-shaped symmetry about the average; the weight distribution of these same men shows a moderate departure from symmetry; while death-rates of a large population at successive age periods assume a U-shaped arrangement, which differs widely in form from either height or weight.

A perfectly symmetrical, bell-shaped distribution has an equal number of items above and below the average or central value. The largest number of cases is concentrated at the mean and the frequencies fall off regularly above and below this value. They are grouped in each class of the lower half in exactly the same manner as in the corresponding class of the upper half of the distribution. When the bell-shaped distribution is represented in the form of a frequency polygon the area of the diagram is divided into two exactly equal and similar parts by the maximum ordinate located at the mean. There are varying degrees of departure from the bell-shaped form, illustrated by the distributions of weight and income.

In a perfectly symmetrical bell-shaped distribution the mean, median, and mode are identical and coincide. This is illustrated by the heights of Japanese soldiers presented in Table 24 and by the frequency curve in Figure 13, page 137.

TABLE 24. HEIGHTS OF JAPANESE SOLDIERS

HEIGHT (inches) m	NUMBER f
(1)	(2)
56	47
57	125
58	316
59	640
60	1,065
61	1,486
62	1,730
63	1,698
64	1,328
65	839
66	442
67	208
68	64
69	12
	10,000

The mean height is 62.24 inches, and the median is 62.26 inches. The mode is not so definitely located as the mean and median, but it is evident that the three forms of average have practically the same value. Whatever slight differences appear between the values may be attributed to *accidental causes.* If the sample were indefinitely increased in size these differences would tend to disappear and the three averages would more and more nearly coincide. This ideal distribution is represented in Figure 13, and it is impossible to distinguish between the values of the mode, median, and mean.

The reader is asked to refer to the five-pound distribution of weights in Table 23 as an example of a moderate degree of departure from the symmetrical bell-shaped form. He should observe the shape of Figure 12 which represents the data of Table 23, and should compare it with Figure 13.

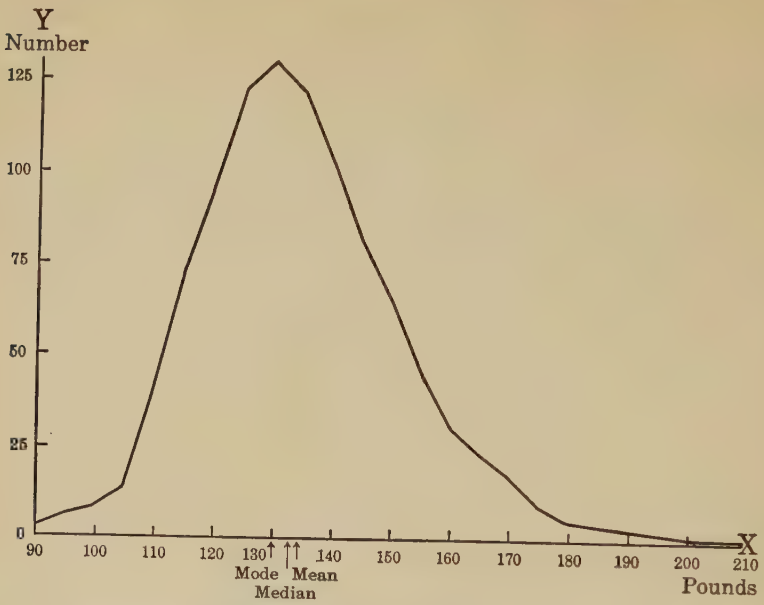


FIG. 12. DISTRIBUTION OF 1000 WEIGHTS — THE FREQUENCIES SMOOTHED BY A TWO-INTERVAL MOVING AVERAGE

The positions of the three averages — Mode, Median, Mean. (Data from Table 23.)

In contrast to Figure 13, in which all forms of average are identical, Figure 12 shows the mode, median, and mean at different values. The mode is about 130 pounds, the median 132.7 pounds and the mean 134.4 pounds. The position of the median should be noted carefully with reference to the other two averages.

As the distribution departs from the symmetrical bell-shaped form the measures of central tendency, the different averages, draw apart. The mode is located at the point of greatest concentration in the distribution, represented by the maximum ordinate, or the peak of the frequency polygon. The mean is drawn farther away from the mode than the median by reason of the influence upon the mean of both the number and size of extreme variants. Both mean and median are drawn in the direction of the margin of the distribution which has the more extreme deviations from the mode. Since it is a position average, the median is influenced by the number of items, greater in one direction than in the other from the mode, but not by the size of the extreme variants.

In distributions which depart only a moderate amount from the bell-shaped symmetrical form, *the median is located about two thirds of the distance from the mode in the direction of the mean.* The position of the median in this case is always between the mode and the mean, but both median and mean may be greater in value than the mode or less in value. In the weight distribution (Figure 12) the extreme variations above the mode predominate and both the median and mean exceed the mode in value. This is the common arrangement of data on weight. The massing point of the distribution is toward the lower end of the range of values. In other distributions the reverse is true. The extreme variations predominate below the mode, and the median and mean are less than the mode. The order is then mode — median — mean, but in the opposite direction from the mode, as indicated in Figure 14, page 138.

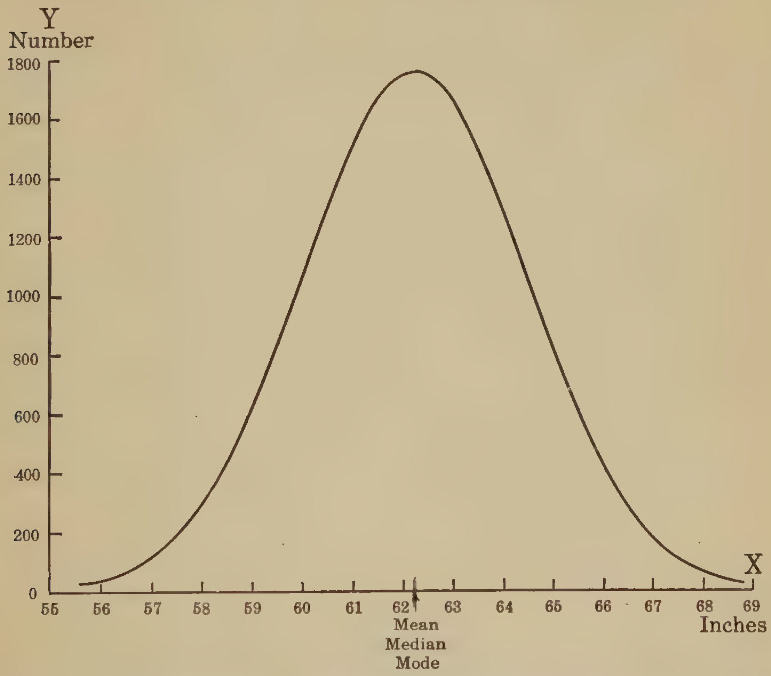


FIG. 13. IDEAL FREQUENCY CURVE FOR THE HEIGHTS OF JAPANESE SOLDIERS
Mean, Median, and Mode identical in position.
(Data from Table 24, Chapter VIII, and Table 47, Chapter XI.)

An empirical rule for locating the mode. Based upon this relationship of the mean and median to the mode, Professor Pearson has given us an empirical rule for computing the mode from the mean and median, which approximates the true mode for moderately asymmetrical distributions:

$$\text{Mode} = \text{Mean} - 3 \ (\text{Mean} - \text{Median})$$

Since both mean and median can be accurately determined, this method is more precise than the crude method of location at the mid-value of the predominant class or by the process of grouping and shifting. Applying this formula to the weight problem, we have

$$\text{Mode} = 134.4 \text{ pounds} - 3\,(134.4 - 132.7) = 129.3 \text{ pounds.}$$

It will be recalled that this value is in accord with our conclusions from

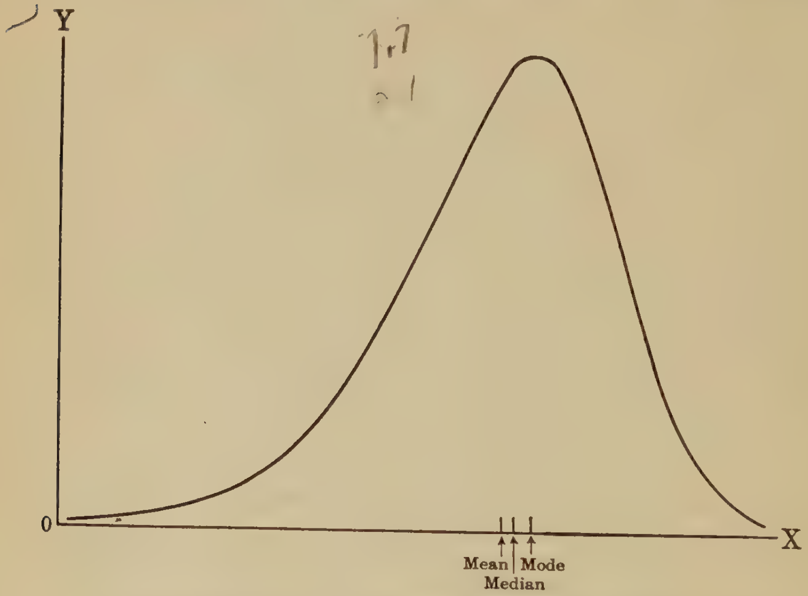


FIG. 14. POSITIONS OF MODE, MEDIAN, AND MEAN IN MODERATELY ASYMMETRICAL DISTRIBUTIONS

(Note the direction of the Mean from the Mode, as compared with Fig. 12.)

the grouping procedure in Table 22. The student must remember to interpret this formula *algebraically*, carefully noting the signs, because sometimes the mean is located above and sometimes below the mode.

The reader may ask at this point why he should use the elaborate method of experimental grouping, illustrated in Table 22, if the simple formula will produce the same result; or under what circumstances the one method will yield more reliable results than the other. In answer to these questions, it may be emphasized that the weight distribution used for illustration is moderately asymmetrical, and that the data are quite regularly distributed about a central value, with a strong tendency to concentrate at or near that value, and with few extreme variations. In such a distribution the formula is reliable and should be used. In general, the larger the number of items the more trustworthy is the formula.

However, in a distribution where there are more extreme variants and where there is a less regular and less concentrated arrangement of the data about some central value, both mean and median are influenced by these conditions. Since in the formula the values of the mean and median are used to compute the mode, the mode when determined in this manner is likely to be less reliable than when determined experimentally by a process of grouping as illustrated in Table 22.

Another method of locating the mode within the class. If the frequencies of classes adjoining that in which the mode appears to fall are symmetrically arranged on either side, it is sufficiently accurate to locate the mode at the mid-value of the class. In actual distributions of social and economic data this rarely happens. Therefore, a method has been used by King and others of weighting the lower or upper half of the class in which the greatest number of items occurs by the frequencies of adjoining classes, thus locating the value of the mode, not at the mid-value, but above or below this point, depending upon the larger frequency in the class above or in the class below. The procedure may be symbolized:

$$\text{Mode} = l + \frac{f_2}{f_2 + f_1}c$$

In the formula l is the lower limit of the class in which the greatest number of items is massed, f_2 is the frequency in the class above, f_1 the frequency in the class below the modal group, and c the size of the class-interval. Sometimes two or even more class-frequencies above are balanced against the frequencies of the same number of classes below the modal class. Applying this procedure to the weight distribution in five-pound intervals, Table 23, and combining the frequencies of two classes above and two below the modal class, 125 to 130 pounds, we have:

$$\text{Mode} = 125 \text{ pounds} + \frac{(125 + 117)}{(125 + 117) + (111 + 81)} \times 5 \text{ pounds} =$$

$$125 \text{ pounds} + 2.8 \text{ pounds} = 127.8 \text{ pounds.}$$

A little experimenting with this formula will make clear that if f_2 and f_1 are equal, the mode will fall at the mid-value of the class; whereas if the frequencies above the modal class predominate the mode will be above the mid-value of the class; and in case the frequencies below the modal class are greater in number, the mode will fall below the mid-value of the class. In other words, the fraction represented by $\frac{f_2}{f_2 + f_1}$ in the formula will be one half if f_2 is equal to f_1, or it will be greater or less than one half as one or the other frequency is larger. This method, while better than that depending upon the mid-value of the modal class, still leaves us the difficulty of changes in the frequencies when the class limits are changed. Adjustment is made only within the class, starting from its established limit. *The method, therefore, is only partly satisfactory.*

The "true mode" in continuous data. The methods illustrated in the preceding pages for locating the mode are really substitutes for the mathematical procedure of fitting a curve to the data, which is the accurate method of locating the "true mode" in a continuous series. Its determination in this manner carries the student beyond the scope of the present discussion.

What we really wish to arrive at, however, in locating the mode in continuous data, *is the mid-value of the class-interval for which the frequency would be a maximum if the intervals could be made indefinitely small, and if at the same time the items could be kept numerous enough so that the frequencies would run smoothly.* The trouble is that we usually have a limited number of cases in an actual distribution, and if we narrow the class-interval beyond a certain point there are too few cases in each class to show any regularity or smoothness in the distribution. The representation of millions of incomes on page 69, classified by $200 and $100 groups, shows how this procedure of smoothing the distribution sometimes can be carried out by narrowing the class-interval. To do this, however, the number of cases must be very large and access to the individual data is necessary. As explained in Chapter V, there are practical objections to too many intervals. Therefore, the mathematical method of fitting a curve to the data is the accurate procedure for securing a *smooth frequency curve. In such a curve the true mode is located at the maximum ordinate.*

Caution should be exercised in the use of exact methods of locating the mode. The reader is asked to review the discussion of continuous and discrete series in Chapter V. In continuous data it is usually possible to measure with greater exactness than is actually recorded. Values may be interpolated between those which are recorded. Provided the number of cases is sufficient the distribution remains smooth and regular as the class-interval is narrowed. *The more refined methods of locating the mode are appropriate only for this type of data.*

In some series measurements are recorded at specific points on the scale and not at other points. Gaps occur because the unit of measurement is not divisible, or because custom determines the unit. Irregular distribution may be due to the character of the data rather than to the paucity of cases. In this type of distribution *care must be exercised not to locate the mode by a process of regrouping or smoothing at a value where items could not occur in practice.* Changing the width of the interval may simply obscure the internal structure of the distribution. In this type of series location of the mode at a specific value within the modal class is justified only when the cases are numerous and when the distribution of the items within that class is known.

APPLICATIONS OF THE MODE

In general the mode is especially useful where the facts are presented graphically because its approximate value and the grouping of the items about it are apparent at a glance. Its use should be confined to series where both frequencies and magnitudes are involved (see Figure 17A), and it should not be used to describe a qualitative category having numerical dominance, as the industry with the largest number of workers. In the study of special aspects of a problem the mode serves to emphasize a particular part of the distribution, as the most common age of graduation from college, the prevailing length of working day, the typical individual contribution in a church collection, the usual size of apartment. The manufacturer of ready-to-wear garments is interested in fitting the largest number of persons. It is not the precise arithmetic mean of various measurements but their mode which gives him the necessary information. The army supply service needs to know the most frequent sizes of hats and shoes. Likewise, in his ordering, the retail dealer in clothing must pay careful attention to the sizes most often demanded and the relative frequency of other sizes grouped about the mode. Otherwise, he will have a number of unwanted sizes on hand.

Wages and hours. The wage earned or the number of hours worked by the most numerous group in a distribution of wages or hours may prove to be the best characterization of wages and hours in a specific occupation. If we are comparing wages at two different periods of time, the mean wage of a group of workers does not show whether improvement is due to leveling up the badly paid, or to rapidly increasing those already well paid. Of course, no single measure can completely reveal the economic condition of the group, but if dependence must be placed on a single value the mode is often the most significant measure.

The mode is also especially useful when estimates must be obtained in the absence of individual wage or employment records. It is the average easiest to estimate. The most frequent value is impressed upon the mind and can be easily obtained by direct questioning.

Prices. The mode is well understood in the field of price statistics.

The most frequently recurrent price, "the usual price," may be far more representative than the mean price, if erratic fluctuations enter into the series. For instance, in January, when ready funds are at a premium, the call money rate of interest may range very high for a few transactions. To take a mean of all call loan rates for the month would not be typical of the prevailing rate for the large mass of transactions of that period. The mode would furnish a far truer picture.

Population statistics. In *anthropometric measurements* the mode often constitutes as true a descriptive value for the series as the mean or median, because of the very symmetrical distribution of the items about the central value; and it is much more quickly approximated, although not as definite in position.

In population statistics the "normal age" of the living population and the "normal age of marriage" are expressions of the most frequent age — the mode. In the analysis of mortality data the same concept is used for characterizing the "most probable time of death." If we start with 100,000 births and calculate from mortality tables the percentage of this number dying at successive ages, the distribution of the deaths will show two modes, one in infancy and the other in old age.

It is clear from Table 25 that if we wish to state in a single figure for each the usual age of marriage for brides and for grooms in Rhode Island

TABLE 25. THE MODAL AGE OF MARRIAGE IN RHODE ISLAND[a]

Age (years) (1)	Grooms *f* (2)	Brides *f* (3)
Under 20	148	962
20 and under 25	2,049	2,327
25 " " 30	1,574	1,129
30 " " 35	768	502
35 " " 40	386	272
40 " " 45	218	131
45 " " 50	124	65
50 " " 55	76	35
55 " " 60	52	22
60 " " 65	28	12
65 " " 70	23	4
70 " " 75	11	1
75 " " 80	4	
80 " " 85	1	
	5,462	5,462

[a] Fifty-seventh Registration Report, Rhode Island, 1909, p. 154.

in 1909, the means will not serve our purpose, because they are too far above the points of massing in the series. The numbers at the higher ages exercise an influence on the mean to draw it away from the point of concentration. The use of the median for this purpose is open to the same objection, especially in describing the ages of the grooms. The mode expresses "the normal age of marriage."

However, if we wish to compare the average age at marriage for all grooms and for all brides, the crude mode, 22.5 years, is the same for both and does not bring out the real difference in age which undoubtedly exists.

The mode in the analysis of industrial accident data. Let us suppose that an investigation is undertaken to ascertain the effect of fatigue upon the number of industrial accidents. The injured workers are classified according to the number of hours worked before the accident, as one hour, two hours, and so on. In Figure 15 these facts are presented graphically.

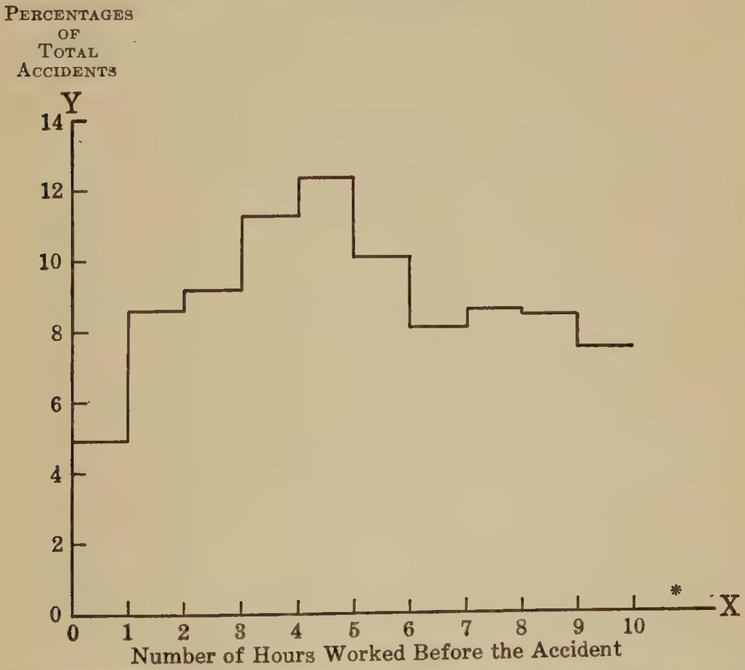


FIG. 15. DISTRIBUTION OF PROPORTIONS OF TOTAL ACCIDENTS OCCURRING AFTER SPECIFIED PERIODS OF WORK

*There is an undistributed group of 10.7 per cent working more than ten hours, indicated on the right of the diagram. (Data from Bulletin 92, United States Bureau of Labor Statistics, p. 49. Total accidents 79,791 = 100 per cent.)

The maximum proportion occurs between the fourth and fifth hours of work, a steady increase having taken place up to this point. One might expect an increase to a maximum at the close of the working period on the ground that increasing fatigue causes accidents, but this increase does not appear. On the contrary, the proportion from the seventh to the eighth hour of work is about the same as that from the first to the second hour of work, and the proportion from the ninth to the tenth hour is distinctly less. The position of the greatest frequency, the mode, is during the fifth hour. This suggests that speed of operation, hourly output if possible, should be compared with the frequency of accidents. During the early hours of the work period the worker probably increases his speed of operation. This activity is accompanied by increase of fatigue and more frequent exposure to dangers. Without attempting here an explanation of the complex problem of causation, it is evident that the mode, as presented in the diagram, is a suggestive measure in the analysis of the data on industrial accidents.

CAUTION IN THE USE OF THE MODE

As we have discussed it in the preceding pages, *the mode is a value, not a qualitative category* as the name of a State, a newspaper, an occupation, a point of time. The mode, furthermore, is *a value which recurs most often in a series of quantitative measures,* not a State, for example, New York, which has the largest population, or a newspaper with the greatest circulation, or an industry with the greatest number of workers, or the hour of the day when the subway traffic is at its height. *The mode describes the usual or normal happening,* the value repeated most often, not the exceptional as in the examples enumerated.

The mode in time series. When presented graphically the mode in a frequency distribution is the value which is represented by the maximum ordinate, the highest point of the frequency curve. In a time series, however, the peak of the curve shows something different from the usual or normal value. It shows rather departure from the normal, and could not be termed modal. In such a series the quantity used to characterize a specific point of time is not a frequency in the sense we have been using the term, although it may appear so at first glance. It is a magnitude which occurs once at the specific time or repeatedly at various times. The divisions of time in the series are for convenience in locating the successive magnitudes. In other words, specific hours, days, months and years at which magnitudes are located are not frequency distributions. Examples will illustrate some possible misuses of the mode.

Figure 16 represents a time series in which are revealed decided *peaks* and *depressions*. But the rectangles plotted vertically from the base line do not represent frequencies in the sense previously used in this chapter, but the volume of traffic at specific points of time. To be significant to traffic officials these points of time, properly designated, do not have to be arranged in sequence of hours, as the class-intervals of a frequency distribution do, although the convenient order of time series is in sequence. The important fact for the operating department to know is that at the period 7 to 9 A.M. and 5 to 7 P.M. the load is at maximum volume. Furthermore, these hours cannot be called modal or normal, in fact their traffic is the opposite of modal. They are the abnormai traffic hours which give the traveling public and the traffic officials so much discomfort and inconvenience. They even constitute a menace by their very unusual character. The normal traffic is represented by the relatively level parts of the diagram, the periods of greatest uniformity. This is a very different situation from that shown in a frequency diagram.

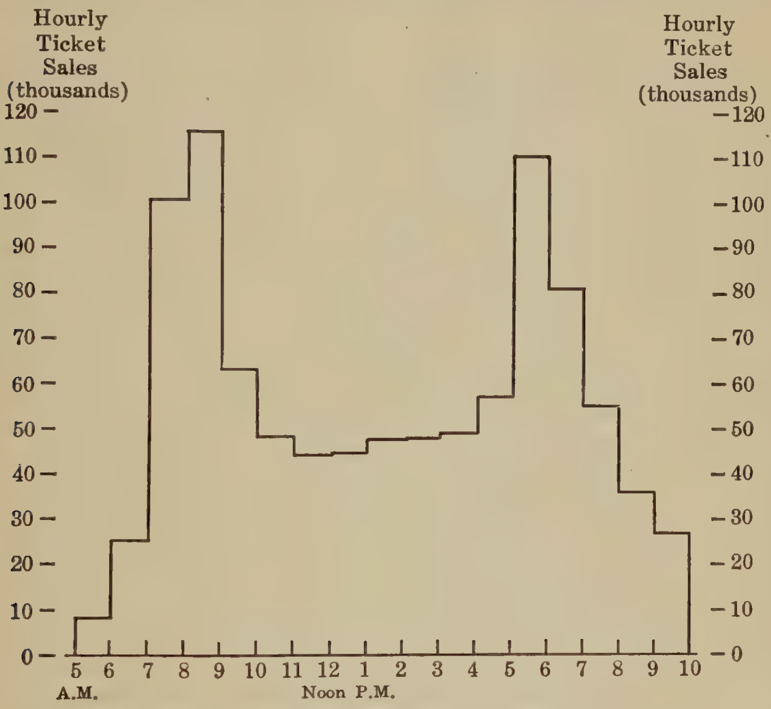


FIG. 16. TRAFFIC ON NEW YORK CITY SUBWAY LINES, JANUARY 22, 1914
(Data from Annual Report, Public Service Commission, First District, 1914, vol. II, pp. 96–97.)

It would be possible to transform the time series of hourly ticket sales into a frequency distribution from the lowest number of tickets sold to the highest. In doing this the time sequence of the hourly sales would be lost. Since the hours of the day when traffic is light or heavy must be identified in order that the traffic department may distribute the cars to best advantage, no useful purpose would be served by changing the form of the series.

If the average rate of interest for call money for each day during the month of January is recorded, great differences may appear between the rates for the first few days of the month and the prevailing rates for the remainder of the month. Due to demand for ready funds at the beginning of the year, interest may be very high for a few days and then fall to normal and continue evenly during the remainder of the month. These high rates must not be termed modal, although a curve showing them graphically for successive days reveals a peak during the early period. These high rates are *abnormal.* They are not frequencies but magnitudes. The frequencies are the number of transactions at a given rate of interest regardless of time. The lower and comparatively level curve for the remainder of the month shows the usual rate of interest.

The modal rate of interest becomes important in this situation because the mean, as explained before, is influenced by the high values and cannot be used with accuracy. Even if the high rates are weighted by the number of transactions, the mean of all the values is still unrepresentative. It is possible to construct a genuine frequency distribution in which the maximum frequency of transactions indicates the prevailing rate of interest. For this purpose the rates of interest considered as magnitudes are arranged in order of size regardless of time, and the number of transactions in each class is tabulated. From this classification the normal rate during the month is at once apparent.

From these illustrations it is evident that *a peak in a time series curve indicates the unusual, the abnormal, which is the exact opposite to the significance of the maximum ordinate in a frequency curve.* The normal in a time series diagram is indicated by the tendency of the curve to run parallel to the base line along which the units of time are recorded, or to show a long time trend upward or downward. A time series must be transformed into a genuine frequency distribution before applying the term mode, but this transformation destroys the time character of the series. It will be more satisfactory to use another term to indicate the high or low positions on the time curve, such as peak and depression. This matter will be further discussed in a later chapter devoted to the treatment of time series (XIII).

The mode in geographic series. The Bureau of the Census publishes periodically the annual per capita expenditures for schools for each city of the United States having 100,000 or more population. The data in Table 26 were taken from the report for 1919.[1]

TABLE 26. PER CAPITA EXPENDITURES FOR SCHOOLS, 1919

New York City	$8.19	St. Louis	$ 6.97
Chicago	6.57	Boston	9.11
Philadelphia	5.35	Baltimore	3.74
Detroit	6.58	Pittsburgh	8.14
Cleveland	7.49	Los Angeles	10.88

Sixty-six cities of 100,000 or more population are recorded in this report in the manner illustrated. *In a geographic series, as in a time series, the quantities placed opposite the locations are not frequencies but magnitudes.* Los Angeles and Boston show the largest per capita expenditures in the table, but they cannot be termed modal. They are exceptional in respect to their per capita expenditures for schools. Furthermore, it will be noted that the names of cities are *qualitative categories* which may be arranged in any convenient order, for example, alphabetically, by sections of the country, or by the amounts expended.

This geographic series of magnitudes which is typical of many other similar series of data may be readily transformed into a frequency distribution.

TABLE 27. CLASSIFIED PER CAPITA EXPENDITURES FOR SCHOOLS, 1919

PER CAPITA EXPENDITURES	NUMBER OF CITIES f
$3 and under $4	6
4 " " 5	5
5 " " 6	13
6 " " 7	15
7 " " 8	11
8 " " 9	10
9 " " 10	4
10 " " 11	2
	66

The per capita expenditures are classified and grouped in order of magnitude and the number of cities expending each given amount is recorded as the frequency in Table 27. It is now possible to summarize the data in terms of the modal expenditure. The most frequent expenditure among the 66 cities is between $6 and $7 per capita.

[1] *Financial Statistics of Cities*, 1919, United States Bureau of the Census, p. 205.

SUMMARY

The possible confusion in the use of the term mode, as previously defined in this chapter, may be avoided if the character of the mode as a *value* is constantly maintained. In graphic representations of data one axis of the diagram, the vertical, represents either frequencies, as in the genuine frequency distribution, or magnitudes located in time or space. The other axis of the diagram, the horizontal, represents quantities, for example, the class-intervals in a frequency table, or points of time, or qualitative categories such as the names of industries or cities. These differences in series of data are shown in Figure 17 A, B, and C.

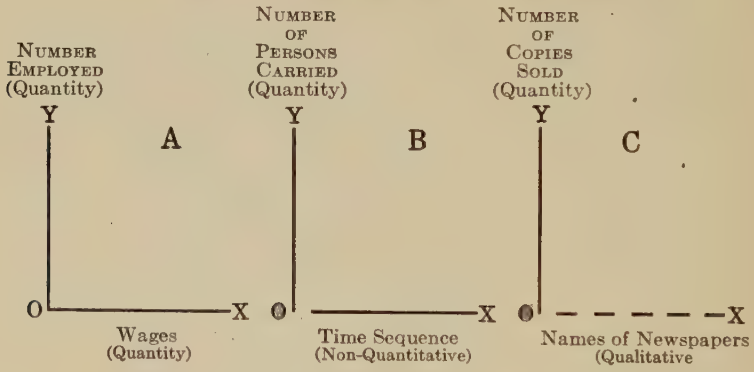

Number Employed (Quantity)
Y
A
O
X
Wages (Quantity)
Number of Persons Carried (Quantity)
Y
B
O
X
Time Sequence (Non-Quantitative)
Number of Copies Sold (Quantity)
Y
C
O
X
Names of Newspapers (Qualitative

Fig. 17. Approved Use of the Mode

In Diagrams B and C the units or designations in the horizontal direction OX are not quantitative. In non-quantitative series the recorded values are not frequencies in the technical sense but magnitudes placed opposite specific categories to describe them. In Diagram B this becomes clearer if the reader will note that the units are not amounts of time (1, 2, 3 hours), but rather sequence or location in time, which is a very different conception. Holding, therefore, to our definition of the mode as a *value concept*, and seeking to locate it along the axis OX in the various diagrams, we find no value which can be called modal except in A where both axes are quantitative. To data such as those in B and C the term mode is not applicable, unless the form of classification is changed from a time-quantity or a quality-quantity basis to a quantity-quantity arrangement.

READINGS

Bowley, A. L., *The Elements of Statistics*, 4th ed., part I, chap. 5.
Rugg, H. O., *Statistical Methods Applied to Education*, chap. 5.
Mills, F. C., *Statistical Methods Applied to Economics and Business*, chap. 4.
King, W. I., *Elements of Statistical Method*, chap. 12.
Jerome, Harry, *Statistical Method*, chap. 7.
Secrist, Horace, *An Introduction to Statistical Methods*, chap. 8.
Yule, G. U., *An Introduction to the Theory of Statistics*, 6th ed., chap. 7.
Zizek, Franz, *Statistical Averages*, tr. by Warren M. Persons, part II, chap. 5 ("Applications of the Mode").
Jones, D. C., *A First Course in Statistics*, chap. 4.
Kelley, Truman L., *Statistical Method*, chap. 3.

[Details as to publisher and date of publication of the readings and references given at the close of each chapter are found in the alphabetical list in Appendix A.]

CHAPTER IX

VARIATION AND ITS MEASUREMENT

One of the purposes of an average is to establish a norm from which to measure the dispersion or variability of individual values. Statistics deals with mass phenomena and in this respect differs from the observation of an isolated case. A single measurement or the record of a few cases gives no basis for judging the significance of the particular value. It is not enough to know that one value differs from another by a specific amount. But after a typical value has been established from sufficiently numerous observations the judgment concerning the individual case may be approached from a new viewpoint, the vantage ground of a knowledge of the behavior of an entire group of similar phenomena with reference to some measurable characteristic.

The importance of variability. For example, from the records of the earnings of individual workers in a trade it is not possible to say with any confidence whether their wages are high or low. These very terms imply an observation wide enough to establish the normal earnings of a representative group in the trade. Individual earnings are high or low with reference to what? Similarly, standards of living of particular families in a group cannot be characterized as abnormal until a normal standard has been established, either from the investigation of family budgets in sufficient numbers to represent the group or by the formulation of an ideal standard of living.

How shall a piece-rate wage be established in a given process of manufacture? What is the normal day's work measured in amount of output? The rate per piece should be established on the basis of average output. Average output is a matter of experience with a representative group of workers. If the most rapid worker's output is used as a basis for fixing the rate of pay, the other workers suffer at the hands of this pace-maker, who is not the type but the extreme variant. Unions often attempt to limit the output of the rapid workers, in the fear that the piece-rate will be gauged according to their output and that the earnings of the other workers will not meet their needs. The effort of the union in such a situation is to require all workers to conform more nearly to the typical output.

The manufacturer is not disturbed so much by a high level of prices of raw materials as he is by fluctuations in prices. He is interested in ways of stabilizing these prices, which means reducing the variations from normal. He wishes to be able to estimate his probable costs with the greatest precision.

The average weight of a large number of healthy babies at each week of the first two years of life is made the basis of judgment concerning the condition of a particular child during its infancy. The terms "overweight" or "underweight," at a given age, mean nothing except with reference to some such standard established from many cases. Knowing how the weights of many babies are distributed about the average and given the weight of a particular child, it is possible to characterize the latter as normal or abnormal.

The death-rate from diphtheria has declined very rapidly in recent years due largely to the use of antitoxin. Progressive control over this disease is indicated not only by the downward trend of the death-rate, but also by the reduction in the fluctuations of the annual rate about this trend, as shown in Figure 18.

Death-rates for any specific cause, as tuberculosis, vary widely in dif-

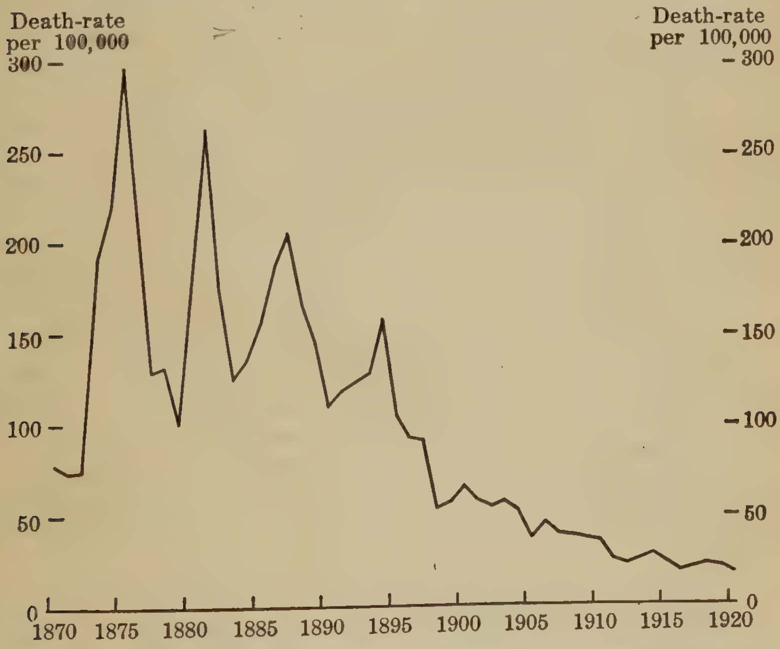


Fig. 18. Annual Death-Rate from Diphtheria and Croup per 100,000 Population, New York City, 1870–1920

(Data from Annual Reports of New York City Department of Health. Before 1898 the rate is for Manhattan and Bronx; for that year and following the rate is for the Greater City.)

ferent sections of a large city. The local rate compared with the general or average rate for the entire city reveals differences in health conditions. Explanation of these variations offers a challenge to health officials. The reduction of these variations constitutes a test of the efficiency of health administration. Why should the tuberculosis death-rate in some areas be under 75 per 100,000 population while in other areas it rises to 400 or over per 100,000? The measurement of variation and its explanation in a geographic series is of fundamental importance.

Degree of homogeneity or similarity in a series. Having established the typical or representative measure by calculating an average, the problem still remains of determining how the individual items are distributed about this central value. In other words, we are still dealing with mass data and we seek a further description of the series to supplement the average. *The average is a point on the scale of measurement, while variability is expressed in some unit of distance on the scale within which is located a known proportion of the items.* We are already familiar with this concept from our discussion of the median and quartiles. Between the two quartile values 50 per cent of the cases are always located. The interquartile range, therefore, is one measure of variability or homogeneity.

In the medical inspection of school children many individuals are described as abnormal in weight. This characteristic is one symptom of malnutrition. What is meant by abnormal? One might answer that the term means that the child of a given age and height falls below the average weight of a large number of healthy children of that same age and height. However, this answer is not satisfactory since all healthy children do not weigh the same amount. Some are above and some below the average weight. Why, then, call the child whose weight is below the average abnormal? How much may a child's weight fall below the average and still not be characterized as abnormal? What proportion of the individual weights of healthy children are included within a range of ten per cent above the average weight and ten per cent below? To determine this fact is an experimental and statistical problem. It is evident that there is a *zone of variation* above and below the average which describes the usual condition of a large proportion of healthy children, and that *the average alone is an inadequate descriptive measure in interpreting the weight of school children.*

It is frequently necessary to compare two or more series of data, for example, the wages of similar groups of workers in several different factories. It is of great importance to know that the wage items in one group are closely massed about the average, while in another group they are relatively much more numerous in the low paid and high paid classes. The mean wage may be about the same in the two distributions and yet the interpretation of the respective wage conditions may be utterly different.

Before inferences are drawn from measurements it is essential to arrange them in some kind of orderly sequence and to examine them carefully as a whole. *Nothing short of the entire distribution is a complete measure of a variable characteristic.* Caution in the use of the average compels the scientific investigator to examine in detail the distribution of the items about the central value. He hopes to discover whether the average employed is really a typical value — as typical as can be found.

application

MEASUREMENT OF ABSOLUTE VARIABILITY

The problem of measuring variability is twofold. First, a multiplicity of individual deviations from the average are to be reduced to a single value, similar to the average itself, representing and describing the entire series. Then, this absolute amount of variation, expressed in whatever unit is being employed in the measurements, must be related to the size of the average from which it is measured and expressed in per cent as *relative variability.*

There are several measures of absolute variability, each covering different unit distances on the scale within which certain proportions of the items are located:

1. The entire range from the highest to the lowest value of the distribution.
2. The semi-interquartile range.
3. The average deviation.
4. The standard deviation.

The range as a measure of variation. If only the lowest wage and the highest are stated in describing the distribution of wage items about the average, the meaning is clear and all the items are included between the given values on the scale. *While this is the simplest measure of variability it is also the least informing.* It gives no idea of the nature of the distribution within these extreme limits. It is very unstable since by cutting off a single wage item at either end of the scale or adding one the range may be entirely changed. It fails to characterize in a useful manner the series as a whole if stated alone, and ignores the degree of concentration almost entirely. It offers no basis for judging the typical character of the average itself.

The semi-interquartile range (Q). In the chapter on the Median the interquartile range has been already discussed as a measure of likeness or difference between measurements. It was pointed out that this range of value on the scale includes 50 per cent of the items, the central half of every distribution. For example, on page 121, it was explained that the central half, 500, of the weight items covered a range of about 22 pounds, while the entire range for the 1000 items was 120 pounds.

In reality the interquartile range is not measured from any central value, as are the other measures of variability to be discussed later. It will be recalled that both quartiles are located by counting the items from the lower or from the upper limit of the array, leaving 50 per cent of the items between the quartiles. In other words, the interquartile range is a distance on the scale such that if we choose at random any single item of the entire array the chances are even that the value of this item will fall between the two quartiles or outside these values.

In a perfectly symmetrical bell-shaped distribution the two quartiles are equidistant from the median, but this is not true for an asymmetrical distribution. In explanation of this difference it must be remembered that the median and quartiles do not divide the range of values into four equal parts, *but rather the number of items, the area of the frequency polygon.* If the structure of the distribution on one side of the median differs from that on the other side, then the range of value on the scale which includes 25 per cent of the cases above the median will differ from that which includes 25 per cent below the median.

In order to treat the interquartile distance on the scale as if it were measured from some central value it has become the practice to take one half the interquartile range as a measure of dispersion about the central value. This is equivalent to measuring from the median to either quartile in a symmetrical distribution. Therefore, the semi-interquartile range equals $\frac{Q_3 - Q_1}{2}$ and includes exactly 25 per cent of the items if the distribution is symmetrical.

In the weight problem with the five-pound grouping the semi-interquartile range equals

$$\frac{144.5 \text{ pounds} - 122.8 \text{ pounds}}{2} = 10.9 \text{ pounds.}$$

This measure of variability is simple in meaning and easily computed. However, it has the limitations which are characteristic of the median as a form of average. In certain forms of distributions the quartiles, like the medians, become indeterminate or unrepresentative. In any case only a part of the items are taken into consideration and the rest ignored.

The average deviation (A.D.). In contrast to the quartile deviation and the range, both the average deviation and the standard deviation are measured from one or another of the measures of central tendency, the mean, median, or mode. All the individual items are given consideration in computing these measures, instead of merely the central half of the items. The average deviation of a series is the summary value representing all the individual deviations from a selected average. It is obtained by taking the mean of all the individual deviations without regard to signs, that is, without reference to the direction of variation. It would be useless to take the algebraic sum of the deviations, which would always be zero when measured from the mean, and in any case would be very small. *The object is to describe the distribution about the central value by a single quantity.* A multiplicity of individual differences is summarized by averaging. The result characterizes the form of an entire distribution and distinguishes it from other similar series for purposes of comparison and inference.

The proper average from which to measure average deviation. *While deviations may be measured from the median, the mean, or the mode, these three measures are not equally defensible on theoretical grounds.* For the computation of the average deviation the variability should be measured from the median because the sum of the simple deviations from the median is less than the sum of the deviations from any other average.[1] It has been explained already that in a symmetrical bell-shaped distribution it can make no difference which average is employed, and in a moderately asymmetrical series the difference is usually small, as the computation from Table 28 demonstrates.

In the table all values in a given class-interval are assumed to be located at the mid-value of the class, column (2). The deviations, column (3), for which the symbol x is used, are obtained by taking the difference between each mid-value and the mean $4.52 ($2.25 — $4.52 = $2.27). These deviations are multiplied by the frequencies, column (5), and without reference to the signs are summed up and divided by the total items. The same procedure is followed in the last two columns, using the median, $4.39, as the central value from which the deviations are measured.

It will be observed that the difference in the average deviation when calculated from the mean and the median is only about one cent, which is insignificant for practical purposes. It should be noted that the sum of the deviations from the median is smaller in the above example than the sum of the deviations from the mean. However, this computation sug-

[1] For proof refer to G. U. Yule, *An Introduction to the Theory of Statistics*, 6th ed., 1922, pp. 144–45.

Table 28. Average Deviation from Mean and from Median

Wage Class	A				B	
	m	Deviations[a] from mean x	f	fx	Deviations from median x	fx
(1)	(2)	(3)	(4)	(5)	(6)	(7)
$2.00–2.49	$2.25	$2.27	6	$13.62	$2.14	$12.84
2.50–2.99	2.75	1.77	16	28.32	1.64	26.24
3.00–3.49	3.25	1.27	34	43.18	1.14	38.76
3.50–3.99	3.75	.77	61	46.97	.64	39.04
4.00–4.49	4.25	.27	66	17.82	.14	9.24
4.50–4.99	4.75	.23	57	13.11	.36	20.52
5.00–5.49	5.25	.73	37	27.01	.86	31.82
5.50–5.99	5.75	1.23	28	34.44	1.36	38.08
6.00–6.49	6.25	1.73	9	15.57	1.86	16.74
6.50–6.99	6.75	2.23	6	13.38	2.36	14.16
7.00–7.49	7.25	2.73	8	21.84	2.86	22.88
7.50–7.99	7.75	3.23	8	25.84	3.36	26.88
			336	301.10		297.20

Mean = $4.52 Median = $4.39

$$\text{A.D.} = \frac{\Sigma fx}{N} = \frac{301.10}{336} = \$.896 \text{ when measured from the mean.}$$

$$\text{A.D.} = \frac{297.20}{336} = \$.885 \text{ when measured from the median.}$$

[a] The symbol x is used to indicate the difference between the individual measurements or the mid-value of a class and the average from which deviations are measured. In this table m − Mean = x. It will be recalled that d indicates the deviations from the guessed average in steps.

gests that it is proper in practice to use the mean in calculating the average deviation if practical considerations make it desirable.

Computation of the average deviation by a short method. Labor may be saved by a short method. We shall use the same data as for the illustration of the long method. Let us assume the average to be at the middle of the class within which it is actually located. To avoid fractions the deviations should be taken in unit intervals or steps from the assumed average, as was done in the short method of calculating the mean itself.

In Table 29 A the procedure is the same as for the long method except that integer steps from an assumed mean are used instead of actual deviations. The sum of the deviations, 616, must be corrected for the difference between the assumed and the true mean before dividing by N, 336, to obtain the final average deviation.

This correction is obtained in the following manner. The difference between the assumed and the true mean is 23 cents ($4.75 – $4.52), or

TABLE 29. SHORT METHOD FOR COMPUTING THE AVERAGE DEVIATION

WAGE	A			B	
m	*f*	Steps from assumed mean *d*[a]	*fd*	Steps from assumed median *d*[a]	*fd*
(1)	(2)	(3)	(4)	(5)	(6)
$2.25	6	5	30	4	24
2.75	16	4	64	3	48
3.25	34	3	102	2	68
3.75	61	2	122	1	61
4.25	66	1	66	0	
4.75	57	0		1	57
5.25	37	1	37	2	74
5.75	28	2	56	3	84
6.25	9	3	27	4	36
6.75	6	4	24	5	30
7.25	8	5	40	6	48
7.75	8	6	48	7	56
	336		$\Sigma fd = 616$		$\Sigma fd = 586$

Assumed mean	$4.75	Assumed median	$4.25
True mean	4.52	True median	4.39

[a] The *d* is used instead of *x*, as in Table 28, to indicate that the deviations are in intervals.

.46 of a step (the steps being each 50 cents). While the class in which the assumed mean is located ($4.75) was assigned a deviation of zero, its actual deviation was $4.75 – $4.52 = 23 cents, or .46 of a step. Therefore, this class and all above it in value show deviations smaller by .46 steps than they should, because the assumed mean is located above the true mean. *Deviations taken on the side of the assumed average are always less than they should be.* This error affects in this case 57 + 37 + 28 + 9 + 6 + 8 + 8 = 153 deviations. The five lower groups show deviations too large by .46 steps, for the same reason. This error affects 66 + 61 + 34 + 16 + 6 = 183 deviations. It is clear, therefore, that 183 deviations in Table 29 A are too large and 153 are too small by .46 of a step each. The net result is that 30 more deviations are too large than too small (183 – 153 = 30), and by this same amount, .46 of a step. The effect of the assumed mean has been to make the sum of the deviations in column (4), 616, too large by an amount equal to 30 times .46 steps, or 13.80 units of deviation. This is the proper correction for the total of column (4). Therefore,

$$\text{A.D.} = \frac{\Sigma fd - \text{correction}}{N} = \frac{616 - 13.80}{336} = 1.792 \text{ steps}$$

$$1.792 \text{ steps times 50 cents} = \$.896$$

This is the true average deviation, and is exactly the same value as obtained by the long method in Table 28. It will be noted that the average deviation is first obtained in terms of steps, to avoid fractions, and is finally reduced to the units of the problem by multiplying by the size of the class-interval.

In Table 29 B the median is used as the central value. The procedure is similar to that in 29 A except that the assumed median is located below the actual median, which is the reverse of the situation in Table A. The median is assumed at $4.25, the mid-value of the class in which the true median, $4.39, is located. The sum of column (6) is 586 which must be corrected. The difference in this case is 14 cents ($4.39 − $4.25), or .28 of a step. Therefore, this group and all below it in value show deviations too small, a total of 183 items. The seven classes above the median class-interval, 153 items, show deviations too large, each by .28 of a step. In this case 30 more deviations are too small than too large (183 − 153), each by the same amount, .28 of a step. The effect of assuming the median has been to make the sum of the step-deviations, 586, in column (6), too small by an amount equal to .28 times 30, or 8.40 steps, the proper correction.

Adding this to 586 steps we have:

$$\text{A.D.} = \frac{586 + 8.40}{336} = \frac{594.40}{336} = 1.77 \text{ steps}$$

$$1.77 \text{ steps times 50 cents} = \$.885$$

This is exactly the same value as that obtained by the long method in Table 28.

The object in using a short method is to save time and labor without sacrifice of accuracy. If the student is not convinced that he can attain accuracy as well as save time he should use the long method. In Table 30 the two methods are exhibited side by side.

In the short method the correction, 43.32 steps, is obtained as follows: The difference between the true and the assumed mean is 1.9 pounds, or .38 of a step, each step being a five-pound interval. The assumed mean is lower in value than the true mean, and, therefore, the deviations of this class and all below it, 557 items, are too small and the other 443 deviation items are too large. The net result is that 114 more deviations are too small than too large (557 − 443), each by .38 of a step. Therefore, the correction equals 114 times .38, or 43.32 steps. This amount must be added to Σfd, 2601, to secure the true sum of the step deviations.

TABLE 30. COMPUTATION OF AVERAGE DEVIATION OF WEIGHTS

POUNDS	A. LONG METHOD (Mean = 134.4 pounds [a])			B. SHORT METHOD (Assumed mean = 132.5 pounds)	
m	f	Deviations from mean x	fx	Deviation in steps from assumed mean d	fd
(1)	(2)	(3)	(4)	(5)	(6)
92.5	6	41.9	251.4	8	48
97.5	7	36.9	258.3	7	49
102.5	10	31.9	319.0	6	60
107.5	18	26.9	484.2	5	90
112.5	65	21.9	1423.5	4	260
117.5	81	16.9	1368.9	3	243
122.5	111	11.9	1320.9	2	222
127.5	134	6.9	924.6	1	134
132.5	125	1.9	237.5	0	
137.5	117	3.1	362.7	1	117
142.5	85	8.1	688.5	2	170
147.5	75	13.1	982.5	3	225
152.5	54	18.1	977.4	4	216
157.5	35	23.1	808.5	5	175
162.5	25	28.1	702.5	6	150
167.5	21	33.1	695.1	7	147
172.5	13	38.1	495.3	8	104
177.5	5	43.1	215.5	9	45
182.5	5	48.1	240.5	10	50
187.5	4	53.1	212.4	11	44
192.5	2	58.1	116.2	12	24
197.5	1	63.1	63.1	13	13
202.5	0	68.1		14	
207.5	1	73.1	73.1	15	15
	1000		$\Sigma fx = 13221.6$		$\Sigma fd = 2601$

$$\text{Long Method: A.D.} = \frac{\Sigma\, fx}{N} = \frac{13221.6}{1000} = 13.2 \text{ pounds}$$

$$\text{Short Method: A.D.} = \frac{\Sigma\, fd + \text{Correction}}{1000} = \frac{2601 + 43.32}{1000} = \frac{2644.32}{1000} = 2.644 \text{ steps}$$

2.644 steps times 5 pounds = 13.2 pounds

[a] The mean is taken here true to the first decimal to make computation simpler. The more exact mean from the five-pound grouping is 134.445 pounds. If the more exact mean is taken, .38 of a step becomes .39 of a step, which will not affect the result 13.2 pounds, true to one decimal.

The standard deviation (*σ* **or sigma**). In the calculation of the average deviation it was necessary to ignore the signs of the deviations in order to secure a quantity which varies with the dispersion of the values. Squaring the deviations is a simple method of making all signs positive. This method is adopted in computing the standard deviation. This measure of variability is the square root of the mean square of all the individual deviations measured from the mean of the distribution.

The deviations are taken from the mean, rather than from the median or mode, because *it is a characteristic of the mean that the sum of the squares of the deviations from it is a minimum.* The mathematical theory employed in the derivation of refined statistical measures makes use of this principle that the sum of the deviations squared should be a minimum.[1] It is also true that squaring gives relatively more influence to the extreme variations than does the averaging of the simple deviations.

Computation of the standard deviation. In a simple ungrouped series the procedure is merely a matter of taking the difference between each measure and the mean, squaring these differences, summing up the results, dividing by the number of items, and extracting the square root.

$$\sigma = \sqrt{\frac{\Sigma x^2}{N}}$$

In a frequency distribution the items in each class are treated as if they were all located at the mid-value, and the deviation of this value from the mean is squared and multiplied by the frequency. These products are summed up and divided by the number of items which gives the *mean square deviation.* The square root of this quantity is the standard deviation.

$$\sigma = \sqrt{\frac{\Sigma f x^2}{N}}$$

The short method used for calculating the mean on page 98 may be extended to the computation of the standard deviation with a great saving of time and labor, and should be employed by the student from the start. In Table 31 the long and short methods of computation are placed side by side. In the computation the difference between the true and the assumed mean is .389, expressed as a fraction of a unit-step. Columns (6), (7) and (8) are already familiar in the computation of the mean by the short method on page 98. Column (9) is obtained by multiplying (7) and (8), in which d is already a factor, by column (6). In this manner the step deviations (d) are again used as a factor, which squares the deviations, makes all signs positive, and permits the entire computation to be done mentally. The contrast in time and labor between the long and the short method is apparent.

The subtraction of the correction factor. In explanation of the use of the correction factor, c, in the short method, it must be recalled that it is a characteristic of the true mean that the sum of the squares of the devi-

[1] G. U. Yule, *An Introduction to the Theory of Statistics*, 6th ed., 1922, pp. 134–35.

Table 31. Computation of the Standard Deviation by the Long and Short Methods

	A. Long Method (Mean = 134.4 pounds)				B. Short Method (Guessed average = 132.5 pounds)			
Weight m (1)	f (2)	Deviations x (3)	x^2 (4)	fx^2 (5)	Step Deviation d (6)	$-fd$ (7)	$+fd$ (8)	fd^2 (9)
92.5	6	−41.9	1755.61	5 10533.66	− 8	48		384
97.5	7	−36.9	1361.61	9531.27	− 7	49		343
102.5	10	−31.9	1017.61	10176.10	− 6	60		360
107.5	18	−26.9	723.61	13024.98	− 5	90		450
112.5	65	−21.9	479.61	31174.65	− 4	260		1040
117.5	81	−16.9	285.61	23134.41	− 3	243		729
122.5	111	−11.9	141.61	15718.71	− 2	222		444
127.5	134	− 6.9	47.61	6379.74	− 1	134		134
132.5	125	− 1.9	3.61	451.25	0			
137.5	117	+ 3.1	9.61	1124.37	+ 1		117	117
142.5	85	+ 8.1	65.61	5576.85	+ 2		170	340
147.5	75	+13.1	171.61	12870.75	+ 3		225	675
152.5	54	+18.1	327.61	17690.94	+ 4		216	864
157.5	35	+23.1	533.61	18676.35	+ 5		175	875
162.5	25	+28.1	789.61	19740.25	+ 6		150	900
167.5	21	+33.1	1095.61	23007.81	+ 7		147	1029
172.5	13	+38.1	1451.61	18870.93	+ 8		104	832
177.5	5	+43.1	1857.61	9288.05	+ 9		45	405
182.5	5	+48.1	2313.61	11568.05	+10		50	500
187.5	4	+53.1	2819.61	11278.44	+11		44	484
192.5	2	+58.1	3375.61	6751.22	+12		24	288
197.5	1	+63.1	3981.61	3981.61	+13		13	169
202.5	0	+68.1	4637.61		+14			
207.5	1	+73.1	5343.61	5343.61	+15		15	225
	1000			$\Sigma fx^2 = 285894.00$		−1106	+1495	$\Sigma fd^2 = 11587$

A. Long Method: $\sigma = \sqrt{\frac{\Sigma fx^2}{N}} = \sqrt{\frac{285894}{1000}} = 16.9$ pounds

B. Short Method: $\sigma = \sqrt{\frac{\Sigma fd^2}{N} - c^2}$

$$c = \frac{+1495 - 1106}{1000} = \frac{+389}{1000} = +.389 \text{ steps}^{a}$$

$$c^2 = (+.389)^2 = .1513 \text{ steps}$$

$$\text{Therefore, } \sigma = \sqrt{\frac{11587}{1000} - .1513} = \sqrt{11.4357} = 3.38 \text{ steps}$$

3.38 steps times 5 pounds = 16.9 pounds

[a] This value, .389 steps, seems to be in disagreement with the .38 steps on page 158. This is because on that page the mean was taken as 134.4 pounds, true to one decimal, while in the short method here used we are carrying the computation out more exactly. This can always be done in the short method without difficulty in computation. This difference does not affect σ true to one decimal.

ations from it is a minimum. Since the step deviations have been measured from an assumed mean (it does not matter whether this value is above or below the true mean) the sum of the squares in column (9) is too large. This will always be true when the deviations are taken from any value other than the true mean. *It follows that the correction which must be made in the mean square* $\frac{\Sigma fd^2}{N}$, 11.587 steps in Table 31, *must always be subtracted.*

The difference between the true and the assumed mean is .389 of a step. This difference affects the deviations of the entire 1000 items, making some too great and others too small, and by the same amount, .389 of a step. Therefore, since all deviations have been squared, we should square .389 and multiply by 1000 to obtain the total correction for the sum of column (9). This product could be subtracted from 11,587, the sum of column (9), before dividing by 1000 and taking the square root. This would be entirely correct, but would require unnecessary labor, as the following symbols show.

$$\sigma = \sqrt{\frac{\Sigma fd^2 - Nc^2}{N}} = \sqrt{\frac{\Sigma fd^2}{N} - \frac{Nc^2}{N}} = \sqrt{\frac{\Sigma fd^2}{N} - c^2}$$

In the formula as stated N cancels out in the correction factor under the radical and leaves c^2 to be subtracted from the sum of the squares of the step deviations divided by N. Therefore, the shortest computation merely requires the squaring of the difference between the true and the assumed mean *in steps*, and the subtraction of this value from the mean square deviation $(11.587 - (.389)^2 = 11.4357$ steps, the true mean square deviation). The square root of the result gives the true standard deviation *in steps*, which must be reduced to pounds by multiplying by 5 pounds, the class-interval (3.38 steps times 5 pounds = 16.9 pounds).

The class-interval as a fraction. The short method is illustrated further in Table 32, where the class-interval is a fraction of one dollar, 50 cents.

It is suggested that the student compute the standard deviation by the long method for this problem to assure himself as to the saving of time and the accuracy of method. *The short method should always be used where possible and the accuracy of the result should be checked by choosing a second guessed average and making the computation a second time, at least to the point where the final root is extracted. Any guessed average should yield the same result under the last radical.*

The class-interval as a single unit. In Table 33 the class-interval is *one* year. The computations are somewhat abbreviated in a distribution

TABLE 32. STANDARD DEVIATION OF A WAGE DISTRIBUTION

WAGE *m* (1)	*f* (2)	STEPS *d* (3)	−*fd* (4)	+*fd* (5)	*fd*² (6)
$2.25	6	−4	24		96
2.75	16	−3	48		144
3.25	34	−2	68		136
3.75	61	−1	61		61
4.25	66	0			
4.75	57	+1		57	57
5.25	37	+2		74	148
5.75	28	+3		84	252
6.25	9	+4		36	144
6.75	6	+5		30	150
7.25	8	+6		48	288
7.75	8	+7		56	392
	336 N.		−201 Nc²	+385	1868

$$\text{G.A.} = \$4.25, \text{ and } c = \frac{+385 - 201}{336} = +.548 \text{ steps}$$

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - c^2} = \sqrt{\frac{1868}{336} - (.548)^2} = \sqrt{5.2592} = 2.29 \text{ steps}$$

and 2.29 steps times 50 cents = $1.15

TABLE 33. STANDARD DEVIATION OF AGES (AGE TO NEAREST BIRTHDAY)

AGE (years) *m* (1)	*f* (2)	*d* (3)	−*fd* (4)	+*fd* (5)	*fd*² (6)
5	4	−3	12		36
6	9	−2	18		36
7	50	−1	50		50
8	86	0			
9	54	+1		54	54
10	24	+2		48	96
11	13	+3		39	117
12	10	+4		40	160
13	5	+5		25	125
14	1	+6		6	36
	256		−80	+212	710

$$\text{G.A.} = 8 \text{ years, and } c = \frac{+212 - 80}{256} = +.516 \text{ steps}$$

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N-1} - c^2} = \sqrt{\frac{710}{256} - (.516)^2} = \sqrt{2.5071}$$

$$= 1.58 \text{ steps, or years}$$

of this kind. Since the class-interval is *one* year the result in *steps* is the same quantity as the result in *years*. No special computation is needed to transform the steps into the units of the problem.

RELATIONS OF THE THREE ABSOLUTE MEASURES OF DISPERSION

The heights of Japanese soldiers in Table 24, page 135, are used to illustrate how the semi-interquartile range, the average deviation and the standard deviation are related to each other in a *bell-shaped symmetrical distribution*. It was demonstrated that the mode, median, and mean of that distribution of heights were practically identical in value. In the ideal distribution represented in Figure 13, page 137, it was impossible to distinguish between these measures of central tendency.

The absolute measures of variability, with the exception of the range, computed for this same distribution of heights are:

1. Semi-interquartile range (Q) = 1.545 inches.
2. Average deviation (A.D.) = 1.805 inches.
3. Standard deviation (σ) = 2.25 inches.

The student should verify these results by making the computations from the data in Table 24.

In Figure 19 the height data are portrayed by three frequency polygons of equal area. The purpose of the diagrams is to compare the numbers and proportions of the entire 10,000 items included within the limits of the interquartile range (Diagram A); within once the average deviation measured plus and minus from the mean (Diagram B); and within once the standard deviation measured plus and minus from the mean (Diagram C). The cross-hatched area of each diagram represents the cases included within the limits of the corresponding measure of variability as laid off on the horizontal scale of magnitudes.

In Diagram A the shaded portion, one half the entire area, represents fifty per cent of the items. If this distribution were perfectly bell-shaped, the quartiles would be equal distances from the median, which would coincide with the mean. If the quartiles were equidistant from the median we would be able to obtain the median by adding the two quartiles (computed from Table 24) and dividing by 2,

$$\left(\frac{60.71 + 63.80}{2} = 62.255 \text{ inches}\right).$$

The median as calculated from Table 24 is 62.26 inches, differing from the computation just made only in the third decimal place. This indicates a very close approach to the perfectly bell-shaped symmetrical

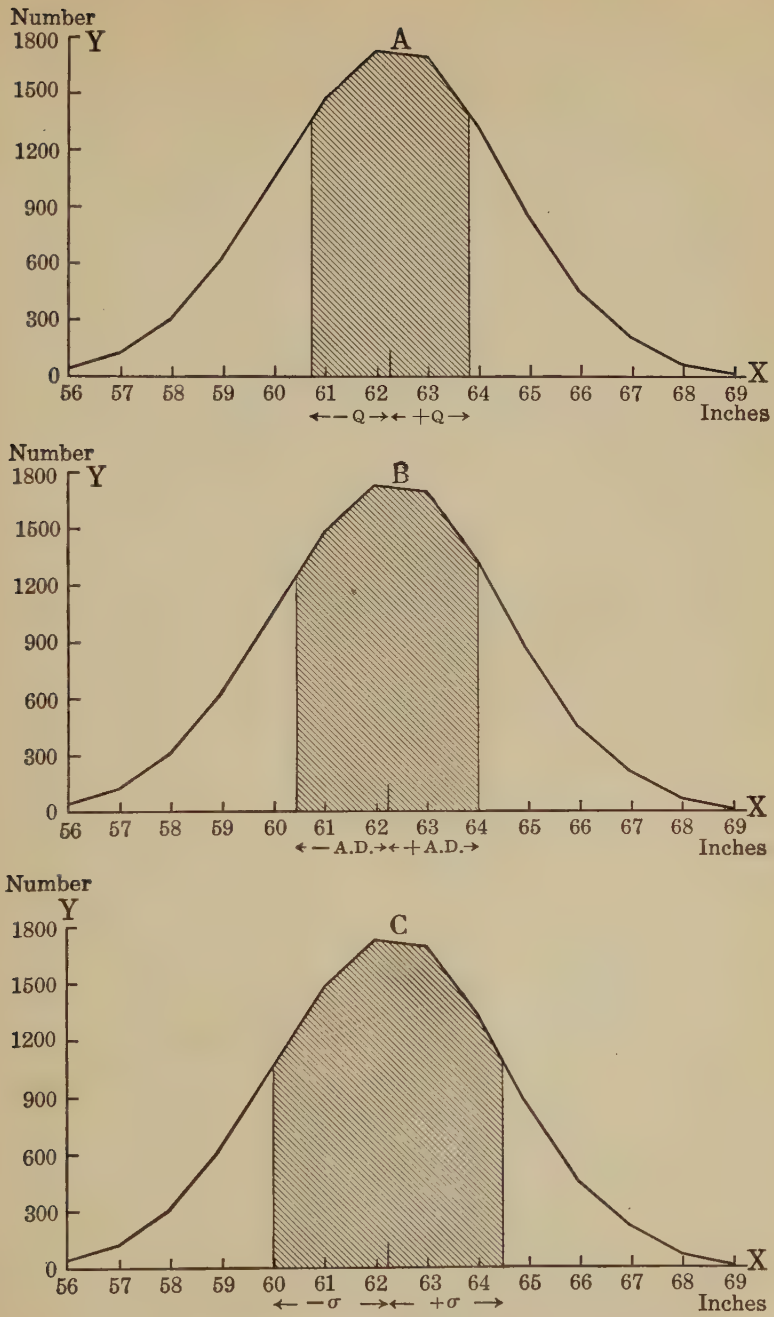


FIG. 19. DISTRIBUTION OF THE HEIGHTS OF JAPANESE SOLDIERS

Representation by shaded areas of the numbers and proportions of total cases included within the interquartile range (A); within the average deviation plus and minus (B); and within the standard deviation plus and minus (C).

distribution. Since the quartiles are so nearly equidistant from the median, the semi-interquartile range, 1.545 inches, is measured plus and minus from the central value in Diagram A, in the same manner as in the case of the average deviation and standard deviation in Diagrams B and C. This procedure is exact only for a perfectly bell-shaped distribution.

In Diagram B the shaded portion represents somewhat more than half the entire number of items distributed over a range of 1.805 inches, the average deviation, above and below the central value (in the perfectly bell-shaped distribution the proportion is 57.5 per cent of the total cases).

In Diagram C the shaded portion represents about two thirds of the 10,000 items distributed over a range of 2.25 inches, the standard deviation, above and below the central value (in the perfectly bell-shaped distribution the proportion is 68.26 per cent of the total cases).

Ratio of one measure of variability to another. *For a perfectly bell-shaped symmetrical distribution constant relations hold between the three absolute measures of variability.* Because the distribution of heights in Figure 19 is not perfectly bell-shaped, the ratios between the measures of variability differ slightly from those obtained from the ideal symmetrical type. Both series of ratios are presented in Table 34.

TABLE 34. RATIOS OF ONE MEASURE OF DISPERSION TO ANOTHER

	IDEAL SYMMETRICAL DISTRIBUTION [a] (1)	HEIGHTS OF 10,000 SOLDIERS (2)	DIFFERENCES [illegible]
σ =	1.2533 times A.D.	1.2465 times A.D.	.0068
σ =	1.4825 “ Q	1.4563 “ Q	.0262
A.D. =	.7979 “ σ	.8022 “ σ	.0043
A.D. =	1.1843 “ Q	1.1683 “ Q	.0160
Q =	.6745 “ σ	.6867 “ σ	.0122
Q =	.8453 “ A.D.	.8560 “ A.D.	.0107

[a] Edward L. Thorndike, *Mental and Social Measurements*, 2d ed., 1913, p. 67.

The differences in the ratios are slight, indicating the close approach of this distribution of heights to the perfect bell-shaped form.

For distributions of the bell-shaped symmetrical type or those which differ from this type only to a moderate degree *the average deviation is about four fifths of the standard deviation, and the semi-interquartile range is about two thirds of the standard deviation.* It is also a useful empirical rule that six times the standard deviation should include ninety-nine per cent or more of the cases in such distributions. These relations furnish a rough check upon the accuracy of the computation.

MEASURES OF RELATIVE VARIABILITY

The computation of averages and measures of variability for a single series of data usually is followed by comparison of these measures with similar ones for other series. For this purpose absolute measures of variation, calculated by any of the methods just explained, have serious limitations. Such measures are directly comparable for different series *only when the averages from which the deviations have been measured are approximately equal in value, and when the units of measurement or estimate are the same.*

There are two difficulties in the comparison of the absolute variability of different series. *First, in distributions where the unit of measurement is the same the size of the average may differ.* For example, the standard deviation of the weekly piece-rate earnings of a group of laborers is $5, measured from an average of $25, while the standard deviation of the annual family incomes of these laborers is $250, measured from an average annual income of $1500. It is clear that for purposes of comparing the variability of the two series the absolute values $5 and $250 have no meaning. The variability measured from $25 could not possibly be as much in absolute amount as when measured from $1500. Yet, if the variability in family incomes is shown to be less than for the earnings of the chief wage earners, it would mean that the family incomes tend to be more alike, that is, more closely grouped about the average. Probably sources of income other than the earnings of the chief wage earner are depended upon to make them so. It may be important to know whether the earnings of individual workers tend to scatter more widely about the central value than do the yearly incomes upon which family standards must depend.

The graphic representation in Figure 20 may throw light upon the problem involved in comparing absolute amounts of variability in different series of data when the same unit of measurement is employed. Three hypothetical frequency distributions are represented by curves A, B, and C. The mean (M_1) is the same for A and B, but the items cluster closer about this central value in A than in B. Therefore, the measure of variability, the standard deviation, for the former is decidedly less than for the latter. It is correct to compare directly the respective standard deviations of the two distributions because the means from which individual deviations are measured are identical. There is a common point of departure in comparing their absolute measures of variability.

It is not so simple to compare measures of variability for B and C.

The diagram indicates that the distributions about the means (M_1 and M_2) are similar. In this case the respective standard deviations are practically identical, but the mean of C is twice that of B. A given amount of variation from M_2 is likely to be less significant than the same amount of variation from M_1. For example, an average variability of five pounds in the weights of a group of adults has much less significance than the same amount for a group of young children.

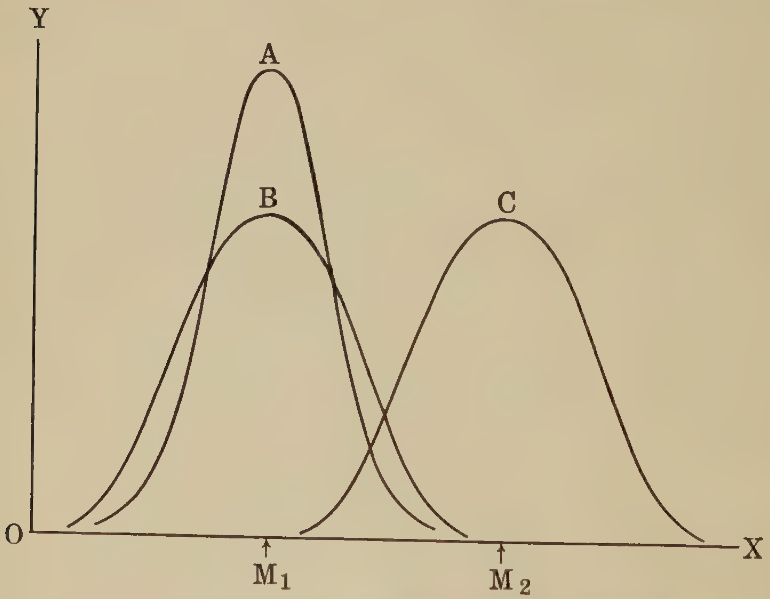


FIG. 20. COMPARISON OF ABSOLUTE AMOUNTS OF VARIABILITY IN DIFFERENT DISTRIBUTIONS

(Importance of the zero origin on the horizontal scale.)

When the averages of two distributions differ in amount the relative significance of a given amount of variability is dependent upon the size of the respective averages from which the individual deviations have been taken.

A second difficulty arises in comparing the variability of distributions in which the units of measurement are not the same, for example weight and height. The standard deviation of the weights of 1000 Freshmen is about 17 pounds and the standard deviation of their heights is about 2½ inches. A comparison of pounds and inches has no meaning. The same difficulty arises in comparing the variation of wages in a specific trade with the variation in years of experience of the workers in that trade.

The importance of the zero origin in graphic representation.[1] In order to emphasize the relation of an absolute measure of variability to the central value from which the deviations have been taken, *it is desirable, if practicable, to represent the zero origin of a frequency distribution on the horizontal scale of magnitudes.* This point was raised for discussion in Chapter V, and in Figure 6 of that chapter the zero origin was shown, as also in Figure 20 of the present chapter. In these diagrams the reader is able to make a more accurate interpretation of the significance of the average amount of dispersion than if the zero origin had been omitted. The size of the average is emphasized by its horizontal distance from zero. The amount of the measure of variability is easily compared with the size of the average by reference to the horizontal scale upon which both are laid off in absolute units.

The coefficient of variation (V). These illustrations have made it clear that some measure of relative variability is needed which takes into account the average from which the deviations are measured, and which reduces different units of measurement to a common basis for purposes of comparison. Professor Pearson has offered a solution by using *as a measure of relative variability* the ratio of the measure of absolute variability (σ) to the mean, expressed as a percentage. This measure is computed according to the following formula:

$$(1) \qquad V = \frac{\sigma \text{ times } 100}{\text{Mean}}$$

Other measures of relative variability have been devised for use when the average deviation or the quartile deviation are employed as measures of absolute variability, and are computed according to the following formulæ:

$$(2) \quad V_{A.D.} = \frac{\text{A.D. times } 100}{\text{Median, Mean, or Mode}}$$

$$(3) \qquad V_Q = \frac{\dfrac{Q_3 - Q_1}{2} \text{ times } 100}{\dfrac{Q_3 + Q_1}{2}} = \frac{Q_3 - Q_1 \text{ times } 100}{Q_3 + Q_1}$$

In formula (2) $V_{A.D.}$ is the coefficient of variation when $A.D.$ is employed as the measure of absolute variability. Usually the $A.D.$ is com-

[1] For a more detailed treatment of this subject the reader is referred to "The Horizontal Zero in Frequency Diagrams," Earle Clark, *Quarterly Publication of the American Statistical Association*, June, 1917, quoted in Secrist's *Readings and Problems in Statistical Methods*, pp. 385–94.

puted by taking the deviations from either the median or the mean but the mode may be used. In formula (3) V_Q is the coefficient of variation when the semi-interquartile range is employed as the measure of absolute variability. All three measures describe the variability as a percentage of the central value, and are independent of the particular units in which the original values are recorded.

In the case of the semi-interquartile range, where the variations are not measured from a central value but between two positions on the scale, the sum of the quartiles divided by two has been used as the denominator of the ratio. In a symmetrical distribution this value is identical with the median.

Sometimes these measures of relative variability are expressed as ratios of the absolute measures to the central values from which the deviations are taken, and are not reduced to the form of percentages, for example, $\frac{\sigma}{\text{Mean}}$ or $\frac{A.D.}{\text{Median}}\cdot$ Instead of being stated as a percentage variation, as 12 per cent, the measure is expressed as a decimal, .12.

Applications of the coefficient of variation. Let us suppose the average weekly earnings of 1000 men in Factory A are $20.50 per week and the standard deviation is $2.20; in Factory B the average earnings are $28.75 per week and the standard deviation is $2.25. The absolute variability in Factory B is slightly greater but comparison of the absolute amounts does not take account of the enlarged possibility of variation by reason of the higher typical wage in the second distribution. Therefore, each standard deviation is reduced to a percentage of the corresponding average.

$$\text{(A)}\quad V = \frac{\$2.20}{\$20.50} \text{ times } 100 = 10.7 \text{ per cent}$$

$$\text{(B)}\quad V = \frac{\$2.25}{\$28.75} \text{ times } 100 = 7.8 \text{ per cent}$$

The *relative variability* of the second factory is shown to be considerably less than the first, whereas its *absolute variability* was greater.[1]

If these data represented conditions in the same factory at two successive periods of time, for example 1913 and 1917, they would indicate not only an increase in the level of wages paid but relatively a closer conformity to the typical wage in 1917. An average variation of $2.25 from $28.75 would not be so significant in terms of the standard of living of the workers as a variation of $2.20 from the lower average wage $20.50.

[1] For an extensive study of absolute and relative measures of variability in wage statistics, see "Employment and Wages," Henry L. Moore, *Political Science Quarterly*, March, 1907.

The closer the average wage approaches to the subsistence level the more serious becomes any variation from it.

These measures of absolute and relative variability are useful in testing the truth of the statement that labor unions *level out* wages by standardizing rates of pay for specific kinds of work, and in this manner stifle individual initiative. It is possible to compare wages in union and non-union shops in the same industry and in the same kinds of work. If the wages of individual workers under union organization are more nearly alike than under non-union conditions, the measures of absolute variability should reveal the fact. Closer concentration about the average wage is indicated by a smaller standard deviation. If union organization has also raised the average wage, then the relative variability under union organization should be decidedly less than under non-union conditions.

Measures of absolute and relative variability will be found useful in gauging the control over disease by the health authorities of a city from decade to decade. Suppose we take the death-rate under one year of age for a large city [1] in 1900 and in 1920. For the same periods we compute the infant death-rates of a number of separate districts of the city. It is possible to compute the standard deviation and the variability for these districts for 1900 and for 1920 in the usual manner. The infant death-rate for the city has declined greatly, as has also the absolute variability of the different sections of the city, as indicated by the standard deviations for 1900 and 1920. But how about the *relative variability* of the districts in 1920 as compared with 1900? A vital question for the health department is whether the differences among the districts have been reduced in relation to the lowered city infant death-rate, that is, whether the relative variability has been made smaller. If so it shows that conditions in the various sections of the city have been made more uniform as far as the control of infant mortality is concerned. If the relative variability remains about the same it shows that the causes of these differences have not been brought under control and it constitutes a challenge to the health authorities to discover and to remove them.

CONCLUSIONS AS TO MEASURES OF DISPERSION

Any summary expression representing the individual deviations may conceal the peculiar form of their actual distribution. The methods described have combined deviations above and below the average with-

[1] This death-rate is not obtained by averaging the rates for the separate districts but by taking the total infant deaths and calculating a new rate for the entire city. This is identical with the *weighted average* of the death-rates of the separate districts. The weights used would be the numbers of infants in the several districts.

out regard to their signs. This is strictly justified only in the case of symmetrical bell-shaped distributions, where the number and size of the deviations above and below the average are the same. There are really two groups of deviations, one positive and the other negative. Most distributions are not perfectly symmetrical. Therefore, computation of the mean deviation may result in a wrong characterization of the distribution if it is very dissimilar in form above and below the average.

In deciding upon the proper measure of absolute variability to use in a specific problem it should be noted that the quartile variation is the simplest to grasp and usually the easiest to compute, if the data are grouped. It proves satisfactory where the median and quartiles are significant as accurate measures of central tendency, especially in fairly symmetrical series of a continuous type where there exists marked concentration of the individual items.

The average deviation and the standard deviation take into account the entire number of items and are influenced by the peculiarities of the outlying variants as well as by those of the central half. The standard deviation has the advantage over other measures from the point of view of arithmetical and algebraic treatment. If the short method is used, the difficulties of computation, for the most part, disappear. Furthermore, the more refined statistical measures, as Pearson's coefficient of correlation discussed in a later chapter, require its use. Experience shows that the average deviation is probably more affected by fluctuations in sampling than the standard deviation. *It is a safe rule to use the standard deviation unless there is a definite reason for choosing another measure of variation.*

The coefficient of variability enables us to compare two or more measures of dispersion and to interpret their relative significance by ranking different distributions according to their tendency to vary from their respective averages. *The values of the coefficients furnish a comparable scale of relative variability.*

MEASURES OF SKEWNESS OR ASYMMETRY

In the measurement of variability all the signs of the deviations have been treated as if positive. *Emphasis was placed upon the amount of variation rather than its direction.* The meaning of a symmetrical as contrasted with an asymmetrical distribution has been explained and illustrated. Certain relations have been shown to exist between the different measures of variability in a symmetrical distribution. These relations become less stable and reliable as the distribution departs from the perfectly symmetrical form. It has been pointed out that most distributions are not perfectly symmetrical, that is, to some extent they show a different structure above and below the central value or average. Moreover, we must compare distributions of varying degrees of symmetry. What measure can be used to describe this departure from the form of perfect symmetry, which we shall call the *skewness* of the distribution? As in the case of the coefficient of variation, this measure in order to be valuable for purposes of comparison must be independent of the units in which the variables are measured.

Our previous discussion of the relative positions of the mode, median, and mean in an asymmetrical distribution will suggest to the student a possible measure for skewness. In the perfectly bell-shaped symmetrical distribution all these measures of central tendency coincide, but in skewed distributions the mean and median are pulled away from the mode in the direction of the skew, or the tail of the curve representing the extreme variants. It will be recalled that the mean is influenced most and that the median moves over a distance about two thirds as great. Therefore, the distance between the mean and median for such distributions is about one third of the entire range between the mode and the mean. *In order to compare the skewness of different distributions, this relation between the mode, median, and mean should be set forth in terms of a common unit of deviation for each series. This common unit of deviation is the standard deviation.*

To meet these requirements Pearson has proposed the following measure:

$$(1)\ \text{Skewness} = \frac{\text{Mean} - \text{Mode}}{\text{Standard Deviation}}$$

This quantity is zero for a bell-shaped symmetrical distribution, and it is positive when the skew is in the direction of the high values, since the mean in this case is always greater than the mode. On the other hand, it is negative when the skew is in the direction of the low values and when in consequence the mean is less than the mode. The mode may be determined by the formula, Mode = Mean − 3 (Mean − Median). Substituting in (1), the numerator becomes 3 (Mean − Median), and

$$(2)\ \text{Skewness} = \frac{3\ (\text{Mean} - \text{Median})}{\text{Standard Deviation}}$$

There is another indication of the existence of skewness, which may have suggested to the reader a second measure. When one quartile differs from the other in its distance from the median, it is clear that the distribution is skewed. To measure the degree of skewness we may take the ratio of this difference to the semi-interquartile range. The following symbols indicate the procedure:

$$(3)\ \text{Skewness} = \frac{(Q_3 - \text{Median}) - (\text{Median} - Q_1)}{\dfrac{Q_3 - Q_1}{2}}$$

$$= \frac{Q_1 + Q_3 - 2\,\text{Median}}{\dfrac{Q_3 - Q_1}{2}}$$

This measure is always zero in a bell-shaped symmetrical distribution and positive when the skew is toward the high values. If we are using the quartile deviation as a measure of variation this measure of skewness is convenient, *although not as sensitive as the first.*

READINGS

Elderton, W. P., and Ethel M., *Primer of Statistics*, chap. 4.
Mills, F. C., *Statistical Methods Applied to Economics and Business*, chap. 5.
Rugg, H. O., *Statistical Methods Applied to Education*, chap. 6.
King, W. I., *Elements of Statistical Method*, chaps. 13 and 14.
Jerome, Harry, *Statistical Method*, chap. 9.
Yule, G. U., *An Introduction to the Theory of Statistics*, 6th ed., chap. 8.
Secrist, Horace, *An Introduction to Statistical Methods*, chap. 11.
—— ——, *Readings and Problems in Statistical Methods*, chap. 9.
Jones, D. C., *A First Course in Statistics*, chaps. 6 and 7.
Bowley, A. L., *Elements of Statistics*, 4th ed., part I, chap. 6.

REFERENCES

Mitchell, Wesley C., Bulletin 284, Bureau of Labor Statistics, *Index Numbers of Wholesale Prices*, pp. 11–23. (Characteristics of distributions of price fluctuations.)
Mitchell, H. H., and Grindley, H. S., Bulletin 165, Agricultural Experiment Station, University of Illinois, July, 1913, *The Element of Uncertainty in the Interpretation of Feeding Experiments*, pp. 472 and 473, and diagrams facing p. 468.
Thorndike, E. L., *An Introduction to the Theory of Mental and Social Measurements*, 2d ed., chaps. 5 and 6. (Excellent graphic representations of variations.)
Pearl, Raymond, *Medical Biometry and Statistics*, chap. 13.
Zizek, Franz, *Statistical Averages*, tr. by Warren M. Persons, part III.
Kelley, Truman L., *Statistical Method*, chap. 4.
Rietz, H. L. (Editor), *Handbook of Mathematical Statistics*, chap. 2.
West, C. J., *Introduction to Mathematical Statistics*, chap. 5.
Davies, G. R., *Introduction to Economic Statistics*, chap. 2.

[Details as to publisher and date of publication of the readings and references given at the close of each chapter are found in the alphabetical list in Appendix A.]

CHAPTER X

INDEX NUMBERS — AN APPLICATION OF STATISTICAL METHODS

INDEX numbers are statistical devices used in measuring relative changes or differences in the magnitude of statistical groups or aggregates of variables.[1] Although their most common applications have been to historical series of commodity prices and wages, their uses have not been confined to these fields. In current use are index numbers of employment, of production and of many other economic series. Sometimes geographical differences are represented by index numbers, with the data for a particular locality as a base for comparison. We shall probably see a rapid and more effective development of the applications of this statistical device in many fields. Therefore, the student should know how to construct and to interpret index numbers.

Changes in real wages — an illustration. Before discussing in detail the methods of construction let us illustrate the use of index numbers in a specific problem. It is desired to show the *change in real wages*[2] of certain classes of railway labor in 1920 as compared with 1913 as a base period. Has the *level of retail prices* of the goods and services for which the laborer spends his wages advanced more rapidly than his money wages or *vice versa?* It will simplify the problem if we use the retail prices of twenty-two representative food commodities, collected and published by the Federal Bureau of Labor Statistics, to represent the changes in the cost of living, since food is by far the largest item of expenditure in the family budget. To make the inquiry complete, changes in the cost of rent, clothing, fuel and light, house furnishing and sundries would have to be included and combined with changes in the cost of food.

In Table 35 the price of each commodity in July, 1920, is compared with its average price for 1913, with the latter as a base, 100 per cent. The percentage change is calculated by dividing each of the 1920 prices in column (3) by the corresponding base-year price in column (2) and multiplying by 100. The result is termed the relative for 1920 on the 1913 base for each commodity. All the relatives are entered in column (4), summed up, and divided by the total number of relatives, 22. *The*

[1] A. A. Young, "Index Numbers," *Handbook of Mathematical Statistics*, p. 181.

[2] Professor Irving Fisher in *The Making of Index Numbers* (published in 1922), p. 368, states: "In Great Britain alone, three million laborers have their wages regulated annually by an index number of retail prices."

TABLE 35. CHANGES IN COST OF LIVING COMPARED WITH CHANGES IN MONEY WAGES

CHANGES IN RETAIL PRICES OF FOOD COMMODITIES

FOOD ITEMS (1)	RETAIL PRICE AVERAGE, 1913 (2)	RETAIL PRICE JULY, 1920 (3)	RELATIVE PRICE IN 1920 (1913 prices = 100) (4)
Sirloin (pound)	$.254	$.487	192
Round (pound)	.223	.450	202
Rib (pound)	.198	.359	181
Chuck (pound)	.160	.286	179
Plate beef (pound)	.121	.191	158
Pork chops (pound)	.210	.437	208
Bacon (pound)	.270	.547	203
Ham (pound)	.269	.597	222
Lard (pound)	.158	.290	184
Hens (pound)	.213	.450	211
Eggs(dozen)	.345	.573	166
Butter (pound)	.383	.679	177
Cheese (pound)	.221	.412	186
Milk (quart)	.089	.167	188
Bread (pound)	.056	.119	213
Flour(pound)	.033	.087	264
Cornmeal (pound)	.030	.070	233
Rice (pound)	.087	.186	214
Potatoes (pound)	.017	.089	524
Sugar (pound)	.055	.265	482
Coffee (pound)	.298	.493	165
Tea (pound)	.544	.746	137
			22)4889
			222

CHANGES IN WAGES OF RAILWAY EMPLOYEES

CLASS OF R.R. LABOR (1)	MONTHLY WAGE 1913 (2)	MONTHLY WAGE 1920 (3)	RELATIVE WAGE IN 1920 (1913 wages = 100) (4)
GROUP A			
Engineers	$178	$260	146
Firemen	107	190	178
Conductors	154	227	147
Brakemen	85	150	176
			4)647
			162
GROUP B			
Baggagemen	87	164	189
Telegraphers	68	140	206
Car repairers	62	146	235
Carpenters	54	130	241
Boiler makers	90	177	197
Machinists	88	170	193
Gang foremen	97	201	207
			7)1468
			210

object is to measure the level of prices of food for 1920 *as compared with* 1913. The average relative 222 is compared with 100, and shows that food prices on the average have increased 122 per cent above the 1913 level, which means that food which cost $1.00 in 1913 cost approximately $2.22 in July, 1920. This was the peak of the upward sweep of retail commodity prices following the World War.

The wage items are treated in a similar manner, grouped in two classes — the better organized brotherhoods, and other skilled groups. The relatives for each group are averaged, showing a 62 per cent increase during the period for the former and a 110 per cent increase for the latter. For neither group have the wages increased as much as the retail prices of food, although for car repairers and carpenters there is a greater increase. Therefore, real wages, as measured by what money wages will buy, have declined, if food prices indicate accurately changes in living costs.

The student is warned that the method here used is not necessarily the best, but the purpose of the index number is set forth in simple terms. *It is a device to measure change, not in one commodity alone but in a level of prices of many commodities.* These various price relatives to be combined are of widely different magnitudes, from 524 and 482 for potatoes and sugar to 137 for tea and 158 for plate beef. The simple average gives equal weight to all the items. Have tea and potatoes and ham equal importance in the family budget? Furthermore, the accuracy of the index number evidently depends upon the accuracy of the original price data and whether they are representative of actual prices paid by housewives. Are these commodities representative in kind and number of the food purchases of ordinary families?

Significant differences will be observed among the various classes of railway labor. The wages of the strongly organized brotherhoods due to their organization were generally higher at the beginning of the period. Their advances were less rapid than the other classes which began at much lower wages in 1913. *This is the reason for dividing into two groups.* The average increase for Group A is much less than for Group B. But in Group A the firemen and brakemen received much more rapid advances than the other two groups. *Should a simple average, which gives equal importance to the slower advances of engineers and conductors, be used to combine the relatives?* What numbers of employees in each class are affected by these different changes in wages? In other words, we face the problem of how to combine in the most accurate manner relatives which vary widely among themselves and which represent items of unequal importance in the combined result. In averaging we recall that the mean is likely to be unduly influenced by extreme variations, such as the relatives for potatoes and sugar. *Consequently, the index number for the food commodities, 222, is probably too high.* This point will be discussed later in the chapter.

REASONS FOR INTRODUCING INDEX NUMBERS AT THIS POINT

It is not intended to offer a complete treatment of the subject in all its technical and controversial aspects. This is an elementary discussion, justified in this connection for the following reasons:

(1) An index number is a summary expression, exemplifying in its construction the *application of the principles of averaging.* It is especially useful in the presentation and interpretation of a time series.

(2) The construction of an index number illustrates certain limitations of the mean in combining relatives and raises the question as to *the proper choice of average to be used.*

(3) The matter of weighting in the calculation of a mean, discussed in a preceding chapter, may be presented in this connection as a practical problem.

(4) The construction of an index number offers an opportunity for a second application of the geometric average.

(5) The gathering of the original data for an index of commodity prices, wholesale or retail, the technique of its construction, and the comparison and interpretation of different series of index numbers give opportunity to apply all the logical steps in a statistical investigation.

(6) An elementary understanding of index numbers is essential for the student of the social sciences that he may be able to read the literature in his field. The student of applied economics and the business executive should understand the construction and uses of index numbers.

A LEVEL OF PRICES

If commodities in the market are represented by numbers, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 and an additional number 11 represents gold, the basic medium of exchange, then, at any particular time, the price of 6, which may be wheat flour, may change, either because of market conditions affecting the demand or supply of wheat flour, or because the value of the money unit, gold, has changed. Over a period of time some commodities are likely to increase in price and others to decline at varying rates. Or, the whole group of commodities increases, as during the war period, but some change at a much more rapid pace than others. We average these price changes or compute an aggregate amount of money which given quantities of the ten commodities cost at different periods. *If the resulting level of prices shows an increase we describe the trend as upward and the purchasing power of the money unit in terms of these commodities moves in the opposite direction.* We are not at present concerned with the variety of causes which produces a change in the level of prices or with the controversial aspects of the quantity theory of money. What we wish to measure with as great exactness as possible is the relative levels of prices at different times or places, or the movement of prices over a period of time. This, index numbers are designed to do.

A VARIETY OF POSSIBLE USES FOR INDEX NUMBERS

It has already been noted that uses are being developed for this statistical device outside the field of prices. Within this field index numbers serve different purposes. To define the specific purpose for which the index is constructed or used is of prime importance because it influences many subsequent decisions. Merely to say, as many have done, that the all-inclusive object is "to measure variations in the exchange value or purchasing power of money" is too vague and indefinite. What does "purchasing power" include? Current index numbers in various countries are constructed, for the most part, from prices of standardized commodities sufficient in number to sample with varying completeness the wholesale markets. They are *general-purpose index numbers.*

The fundamental question arises as to what the prices included in the given index really represent. The movement of what field or type of prices do we seek to describe by an index? Interest may be concentrated upon commodities sold at retail, prices of industrial stocks, prices of raw materials, wages in a specific industry, interest rates — a great variety of goods and services which are bought and sold. Index numbers have been developed recently to measure fluctuations in the physical volume of production as well as changes in the money values. These are *special-purpose index numbers* to aid in understanding specific economic and social problems.

The illustration at the beginning of this chapter represents a familiar use of index numbers to show changes in the cost of living. So far as this index is representative of the changes in the cost of the family budget, it enables us to transform the changes in *money wages* into changes in *real wages in terms of commodities and services.* A board of wage arbitration, in turn, may use these results to decide what change in the money wages now paid should be made in order to protect the standard of living of the worker against the menace of rising prices, or to adjust wages to a declining level of prices in a period of business deflation. The situation would not be fairly reflected by wholesale prices because retail and wholesale prices do not vary either at the same time or to the same degree. *Living costs are in terms of retail prices.*

On the other hand, an index may be used to show the general change in purchasing power of the standard money unit. This is one series of data used for describing the periodic movements in business through phases of prosperity and depression — the business cycle. Such a general-purpose index is that of the United States Bureau of Labor Statistics in which a large number of wholesale commodity prices are combined. The great volume of buying and selling in the business community is done at wholesale before the finished goods are ready for the ultimate consumer. Furthermore, it is clear that such a general commodity index, designed to represent the entire field of transactions, must utilize many more items than a retail index for food. These items must be selected from different types of commodities, representing various stages of production from the raw material to the finished product, whose prices behave in different ways. The object is to represent fairly the various groups of commodities in the entire price situation, and to give to each group its proper influence in the final combined result. The Bureau of Labor Statistics, for this purpose, classifies commodities into nine groups — farm products, food, cloths and clothing, fuel and lighting, metals and metal products, building materials, chemicals and drugs, house furnishings, and miscellaneous. For each of these groups representative commodities are selected, the wholesale prices are collected regularly, and an index is constructed, first for each group and then for all commodities combined — a single general commodity index of wholesale price changes.

Our purpose at this point in the discussion is to caution the student against the conclusion that one index number is as good as another for a specific purpose. *The use to be made of the index is a fundamental consideration, whether we are constructing one for our own use or are selecting one already in use.*

STEPS IN THE CONSTRUCTION OF A PRICE INDEX

The enumeration and brief discussion of the steps in construction indicate the statistical problems involved and suggest methods for their solution. The problems are:

(1) Defining the purpose of the index number.

(2) Deciding the number and kind of commodities and whether retail or wholesale prices are required.

(3) Collecting the actual price data and the facts necessary for weighting.

(4) Deciding the method of combining the items into a single index. If items are averaged, deciding which average to use.

(5) Deciding upon a base period from which to measure changes.

(6) Deciding that equal importance shall be assigned to each commodity or that weights shall be assigned according to relative importance.

NUMBER AND KIND OF COMMODITIES

With the purpose of the index defined, it becomes an easier task to decide the number and kind of commodities required and whether retail or wholesale prices are more appropriate. It is necessary to study the list of commodities included in any ready-made index in order to know what it really measures. One group of price data serves better than another as a "business barometer," while a very different list of commodities is required to reflect changes in the cost of the family market basket.

The number of items depends upon the field of prices to be represented. This is a problem of sampling. Until recently the food index of the Federal Bureau included 22 items, but the number has been increased to 43. The inclusion of the additional items makes less difference in the index than might be supposed, because the original commodities had been carefully selected and were given greater importance by weighting than the added commodities. On the other hand, the general wholesale index of the Bureau includes several times as many items as the present retail index for food, in order to represent adequately the various groups of commodities. The fluctuations in the prices of raw materials and their manufactured forms, for example, pig iron and steel instruments, wheat and bread, are related but not identical. Raw material prices as a rule fluctuate more widely. Likewise, there are characteristic differences in the price changes of mineral products, animal products, and farm crops. The prices of consumers' goods do not behave in the same manner as those of producers' goods, for example, food, clothing and house furnishings as compared with tools and factory equipment.[1]

The practical conclusion is that any index number measuring changes in general commodity prices must include samples from the various groups which behave in different and characteristic ways. It follows also

[1] The reader will find detailed tables and diagrams describing the fluctuations of wholesale prices of single commodities and groups of commodities in Bulletin 320, United States Bureau of Labor Statistics, *Wholesale Prices 1890 to 1921.*

that the general index which combines a larger number of well selected items is likely to be the more trustworthy as a representative of the commodity price level. Index numbers made for specific groups of commodities require correspondingly fewer items. *It is a question of representing adequately the field covered by the samples.*

COLLECTION OF PRICE DATA

One of the most important influences promoting wider use of index numbers and greater confidence in their reliability has been the recent improvement in price records and the extension of the scope of the data. In the field work of collecting prices *accuracy and representativeness are of fundamental importance.* In recording wholesale prices shall we use market quotations, contract prices, or export and import prices? What is the retail price of eggs in New York City? The answer is not so simple as it may at first appear. Even on a given day in the same market, prices of the same commodity differ. Different grades of eggs have different quotations. Careful definition of the kind and grade of commodity is essential. *There is constant danger of regarding things which are really different as alike, and things which are really alike as different.* This difficulty is greater in the collection of retail prices than of wholesale. Many commodities, especially raw materials such as wheat and cotton, have been graded and designated by widely known and accepted terms. For wholesale transactions the market is wide and fairly easily determined, but at retail, prices vary with location even within the same city, with the size of the store, the system of delivery, credit or cash systems, etc. How will the field worker secure price data for a commodity which can be reduced to a single figure for an entire city?

The task is to select for each commodity the quotations which are most typical and in sufficient number to represent the factors which control varying prices in the community. The investigator seeks the most reliable source, the most representative market and the typical grades. Quotations must represent uniform qualities and allowance must be made for cash discounts, premiums, credit and other known factors.[1]

METHODS OF COMBINING PRICE DATA

Two chief methods of combining the price changes of many commodities into a single index are in common use.

(1) *Method of averaging relatives.* This method was illustrated at the

[1] The procedure followed by the United States Bureau of Labor Statistics in collecting price data, retail and wholesale, is described in their Bulletins which can be obtained on request at the office in Washington. It is possible to secure also sets of blank forms used in recording the data and the instructions given to the agents of the Bureau.

beginning of the chapter. No attempt was made to weight the relatives according to the importance of the commodities. The matter of weights will be discussed later.

(2) *Comparison of aggregate values of definite quantities of the selected commodities.* These aggregates are sums expressed in dollars and cents, obtained by multiplying the price of each commodity at the given period by a definite amount of the commodity and adding the separate products. The result is the aggregate value of a bill of goods at a specific time or place. These totals are compared as such without averaging.[1]

A modification of the second method merely adds one more step. It relates the aggregate values to a common base period, the aggregate value of which is called 100 per cent for convenience in comparison. Using the aggregate value of the base period as a divisor, all other values are reduced to relatives or percentages of this base value, as in the first method. The difference between the first method and this modification is that the relatives in the second method are computed at the end of the process for convenience in comparing the aggregate values.

Method of averaging relatives. This was the method used for many years by the Bureau of Labor Statistics, but it was abandoned in 1914 in favor of the method of comparing aggregate values. Some of the advantages of reducing the price changes for each separate commodity to relatives before combining into a single index will appear in the simple illustration of wholesale prices in Tables 36 and 37.

TABLE 36. COMPARISON OF AGGREGATES OF ACTUAL PRICES

COMMODITY (1)	AVERAGE 1913 (2)	JULY 1917 (3)	JULY 1918 (4)	JULY 1919 (5)	JULY 1920 (6)
Wheat (bushel)	$.874	$2.582	$2.170	$2.680	$2.831
Flour (barrel)	4.584	12.750	10.702	12.155	13.669
Sugar (pound)	.043	.075	.074	.088	.191
Hogs (100 pounds)	8.365	15.460	17.720	22.225	14.856
Eggs (dozen)	.226	.318	.374	.406	.423
All commodities	14.092	31.185	31.040	37.554	31.970

From Table 36 it will be observed that the quotations are for different units of each commodity. It is difficult to combine them by simple addition into a single significant index of price change. Sugar increased more than four-fold during the period, but the importance of this change

[1] Bradstreet's index number is an *aggregate of actual wholesale prices* per pound of a large number of representative commodities.

in price is submerged in the larger values of the other units summed up in columns (2) and (6). Bradstreet's index is obtained by first reducing each commodity value to a price per pound and then summing up the results, as in Table 36.

On the other hand, the price changes for each commodity may be reduced to a percentage of some common base called 100. This method avoids the difficulty of the variety of units for which quotations are secured. In Table 37 the 1913 prices of Table 36 are used as divisors for all the other price items and the results are multiplied by 100. Now the items to be combined become relatives.

TABLE 37. COMPARISON OF RELATIVE PRICES

(Average Prices of 1913 equal 100)

COMMODITY	AVERAGE 1913	JULY 1917	JULY 1918	JULY 1919	JULY 1920
(1)	(2)	(3)	(4)	(5)	(6)
Wheat	100	295.4	248.3	306.6	323.9
Flour	100	278.1	233.5	265.2	298.2
Sugar	100	174.4	172.1	204.7	444.2
Hogs	100	184.8	211.8	265.7	177.6
Eggs	100	140.7	165.5	179.6	187.2
Average of relatives	100	214.7	206.2	244.4	286.2

To secure a single index for the price changes of the five commodities we have used the simple average, adding each column and dividing by 5. This method gives equal importance to each relative and leaves the problem of weighting for discussion later.

Shifting the base period for comparison. The investigator may wish to compare all other prices with those of the year 1918. For this purpose the base must be shifted from 1913 (100) to 1918 (100). Frequently a short method has been used to avoid the labor of recomputing the index from the actual prices of the new base. This procedure involves simply dividing the series of index numbers obtained in Table 37 by the index number of the new base year, in this case 206.2, and multiplying each result by 100.

	1913	*1917*	*1918*	*1919*	*1920*
206.2)	100	214.7	206.2	244.4	286.2
	48.5	104.1	100	118.5	138.8

In other words, if 206.2 for 1918, according to the old series of index numbers on the 1913 base, be made 100, the new base, what will the index numbers for the other years become? The result shows the rise in the prices of this restricted group of commodities after the close of the war to a new level 38.8 per cent higher in July, 1920, than in the same month of 1918.

This is a very simple procedure for shifting the base, but the short method gives inconsistent results when the original index numbers have been computed by the method of simple arithmetic averaging of relatives as in this example. The results do not coincide, as a rule, with the index numbers recomputed on the actual prices of 1918 (100). Table 38 uses the 1918 prices from Table 36 as a new base and shows the recomputed relatives.

TABLE 38. COMPARISON OF RELATIVE PRICES COMPUTED ON NEW BASE

(Prices of July, 1918, equal 100)

COMMODITY (1)	AVERAGE 1913 (2)	JULY 1917 (3)	JULY 1918 (4)	JULY 1919 (5)	JULY 1920 (6)
Wheat	40.3	119.0	100	123.5	130.5
Flour	42.8	119.1	100	113.6	127.7
Sugar	58.1	101.4	100	118.9	258.1
Hogs	47.2	87.2	100	125.4	83.8
Eggs	60.4	85.0	100	108.6	113.1
Average of Relatives	49.8	102.3	100	118.0	142.6

The 1920 index number in Table 38 differs by about four points from the result obtained for that year by the short method, and the indexes for the other years differ, but in less degree.[1] The recalculation is usually too laborious even if all the price data were available, which is frequently not the case.

An advantage of the geometric average. The geometric average if used to combine the relatives in Table 37 will produce a series of index numbers in which the base can be shifted with mathematical accuracy by the short method. In Chapter VII the method of computing a geometric average was explained. For any given year in Table 37 the arithmetic mean of the logarithms of the relatives for that year will be the logarithm of the index number. The natural number corresponding to this logarithm will be the required index for that year. Table 39 indicates the method, in which the logarithms given are for the relatives in Table 37.

[1] For more detailed explanation and illustration of why the arithmetic means of relative prices cannot be consistently shifted to another base by the short method, see Bulletin 284, United States Bureau of Labor Statistics, pp. 83–85. This bulletin was prepared by Professor Wesley C. Mitchell and the author acknowledges with appreciation his indebtedness to Professor Mitchell and to the Bureau for much of the material of this chapter. This

Table 39. Combining Relative Prices by the Geometric Mean
(Relatives from Table 37 — Use of Logarithms)

Commodity (1)	Average 1913 (2)	July 1917 (3)	July 1918 (4)	July 1919 (5)	July 1920 (6)
Wheat	2.00000	2.47041	2.39498	2.48657	2.51041
Flour	2.00000	2.44420	2.36829	2.42357	2.47451
Sugar	2.00000	2.24155	2.23578	2.31112	2.64758
Hogs	2.00000	2.26670	2.32593	2.42439	2.24944
Eggs	2.00000	2.14829	2.21880	2.25431	2.27231
Average Logarithm	2.00000	2.31423	2.30876	2.37999	2.43085
Index	100	206.2	203.6	239.9	269.7

The index numbers obtained by the geometric method of combining relatives are all different from those of Table 37, and *each is smaller. It is characteristic of the geometric mean to minimize the influence of extreme variations.* The greatest difference is in the index for 1920, because the dispersion of the relative prices in that year is wider than in any other.

We shall shift the base from 1913 to 1918 by the short method of dividing all the index numbers of Table 39 by the index of the new base year, 203.6, and multiplying each result by 100.

	1913	*1917*	*1918*	*1919*	*1920*
203.6)	100	206.2	203.6	239.9	269.7
	49.1	101.3	100	117.8	132.5

This new series of index numbers should be exactly the same as the indexes obtained by recomputing the relatives from the actual prices on the 1918 base (100), as in Table 38, and combining the relatives by the geometric average. We shall test this by taking the logarithms of the relatives in Table 38 and securing the mean for each year and the resulting indexes. If the results are the same as those obtained by the short method, we shall have confidence in this method of shifting the base, provided the original relatives have been averaged by the geometric mean as in Table 39.

The resulting index numbers are identical with those obtained by the short method. It is important to be able to shift the base period at will and with accuracy by a short method, since the indexes can then be used by those interested in comparisons other than those with the original fixed base. Many consumers of index numbers wish to make different com-

bulletin is the best available introduction to the discussion of price index numbers and the principles of their construction, and is especially valuable for those who want a non-mathematical presentation.

TABLE 40. COMBINING RELATIVE PRICES BY THE GEOMETRIC MEAN
(Relatives from Table 38 — Use of Logarithms)

COMMODITY (1)	AVERAGE 1913 (2)	JULY 1917 (3)	JULY 1918 (4)	JULY 1919 (5)	JULY 1920 (6)
Wheat	1.60531	2.07555	2.00000	2.09167	2.11561
Flour	1.63144	2.07591	2.00000	2.05538	2.10619
Sugar	1.76418	2.00604	2.00000	2.07518	2.41179
Hogs	1.67394	1.94052	2.00000	2.09830	1.92324
Eggs	1.78104	1.92942	2.00000	2.03583	2.05346
Average Logarithm	1.69118	2.00549	2.00000	2.07127	2.12206
Index	49.1	101.3	100	117.8	132.5

parisons from those intended by the original compilers.[1] The very fact that this cannot be easily done when the unweighted arithmetic average is used counts against this method of combining relatives and in favor of the geometric average.

On the other hand, the geometric form of average is comparatively unfamiliar to those who use index numbers, and is somewhat laborious in computation. Is there another method of combining price data which retains the advantage of ease and accuracy in shifting the base, and at the same time avoids the disadvantages of the geometric average?

Method of comparison of aggregate values. This method involves the use of data on both prices and quantities of the commodities. The quantities of the various commodities whose prices are to be combined constitute weights. For example, we may use the actual amounts entering into exchange relations of buying and selling, in constructing a general-purpose index number; or the average amounts consumed in the ordinary family budget, if a retail index is desired for the purpose of measuring changes in the cost of living. *In either case this method involves the comparison of the costs of a definite bill of goods at different times or places.*

The quantities of food commodities consumed by a typical family have been investigated by the Bureau of Labor Statistics and by other agencies through inquiries concerning family expenditures in various sections of the country. The retail index for 22 food commodities, used for illustration at the beginning of this chapter, is now presented in the form used by the Federal Bureau of Labor Statistics, with the exception that the list of commodities has been increased.

Table 41 furnishes aggregate values for 1913 and July, 1920, of a defi-

[1] See *The Labor Market Bulletin*, New York State Industrial Commission (June, 1920), vol. 6, no. 6, p. 5, for illustration of the need for shifting an old series of index numbers to a new base to make two series comparable.

Table 41. Comparison of Aggregate Values at Retail 1913 and 1920
(Aggregate value 1913 equals 100)

Commodity (1)	Average Price 1913 (2)	Quantity Consumed (3)	Aggregate Value (2) × (3) 1913 (4)	Price July 1920 (5)	Aggregate Value (3) × (5) July, 1920 (6)
Sirloin (pound)	$.254	70 lbs.	$17.780	$.487	$34.090
Round (pound)	.223	70 "	15.610	.450	31.500
Rib (pound)	.198	70 "	13.860	.359	25.130
Chuck (pound)	.160	70 "	11.200	.286	20.020
Plate beef (pound)	.121	70 "	8.470	.191	13.370
Pork chops (pound)	.210	114 "	23.940	.437	49.818
Bacon (pound)	.270	55 "	14.850	.547	30.085
Ham (pound)	.269	55 "	14.795	.597	32.835
Lard (pound)	.158	84 "	13.272	.290	24.360
Hens (pound)	.213	68 "	14.484	.450	30.600
Eggs (dozen)	.345	85 doz.	29.325	.573	48.705
Butter (pound)	.383	117 lbs.	44.811	.679	79.443
Cheese (pound)	.221	16 "	3.536	.412	6.592
Milk (quart)	.089	355 qts.	31.595	.167	59.285
Bread (pound)	.056	225 lbs.	12.600	.119	26.775
Flour (pound)	.033	454 "	14.982	.087	39.498
Cornmeal (pound)	.030	227 "	6.810	.070	15.890
Rice (pound)	.087	25 "	2.175	.186	4.650
Potatoes (pound)	.017	882 "	14.994	.089	78.498
Sugar (pound)	.055	269 "	14.795	.265	71.285
Coffee (pound)	.298	47 "	14.006	.493	23.171
Tea (pound)	.544	11 "	5.984	.746	8.206
Aggregate values			$343.874		$753.806
Index			100		219.2

nite bill of food commodities, apportioned according to the average quantities actually consumed in a large number of representative families, column (3). The quantities remain the same for both periods compared. The variables are the prices and the aggregate values. This method has the merit of avoiding the difficulties of averaging since the aggregates are compared directly on any desired base (100), without averaging.

Besides, the base may be shifted accurately by the short method. If we wish to change to the 1920 base (100), it is only necessary to divide by 219.2, the index of the new base, and to multiply each result by 100.

	1913	*1920*
219.2)	100	219.2
	45.6	100

Exactly the same result is obtained by using $753.806 as the new base and calculating the percentage which $343.874 forms of this amount;

that is, by calculating a new index from the original aggregate values ($343.874 ÷ $753.806 times 100 = 45.6).[1]

This method of aggregates has particular merit in the measurement of changes in the cost of living because the housewife is interested in the changes in value of a bill of goods at successive periods. In other words, this index serves a clear and specific purpose and the result is easily understood. *The Federal Bureau of Labor Statistics uses this method at present for both its retail and wholesale index numbers.*

CHOICE OF THE BASE PERIOD

There are two types of index numbers from the point of view of the base period: (1) *the fixed-base index* such as the examples already used in this chapter, and (2) *the chain or link relative index.* The chain index uses each year's or month's price as the base (100) upon which the relative for the following year or month is calculated, and so on year by year or month by month. Likewise, the aggregate values may be secured and each successive aggregate used as a base (100) for the following month or year. By this link relative index only year to year or month to month changes are measured. For some purposes this is the comparison most desired. It is easy to add or drop commodities from the list in such an index because the comparison is over a limited period and, therefore, *the compiler is free to revise his list of items whenever it seems desirable.* On the other hand, it is difficult to preserve the representative character of the items in a fixed-base index which extends over a long period of time.

Relating chain indexes to a fixed base. The link relatives in a chain index may all be related to a fixed base, forming a continuous series. To illustrate the method we shall take the prices of crude oil in the Pennsylvania field from Table 43 A, page 192, and compute link relatives to form a chain index, using the price of each year as the base (100) for the following year. Then we shall relate all the relatives to the base year, 1913, by a series of multiplications. The results are shown in Table 42.

The first link relative (76.7) in column (3) is already related to the fixed base (100 for 1913). The second link relative (82.7) is chained back to the same base by multiplying it by the first relative, 76.7, to which it has been linked. The third relative (160.4) is related to the base, 1913, by multiplying it by the product already obtained in the previous chaining process (63.4), and the last relative (129.8) is multiplied in turn by the product obtained from the previous chaining process (101.7). This procedure can be continued for a longer series until each

[1] For the algebraic proof of this procedure, see Bulletin 284, United States Bureau of Labor Statistics, p. 79, footnote.

TABLE 42. LINK RELATIVES IN A CHAIN INDEX RELATED TO A FIXED BASE

YEAR	PRICES OF CRUDE OIL (barrel) (1)	RELATIVES ON FIXED BASE 1913 = 100 (2)	LINK RELATIVES, EACH YEAR = 100 IN SUCCESSION (3)	LINK RELATIVES CHAINED TO FIXED BASE, 1913 = 100 (4)
1913	$2.463	100.0	—	100.0
1914	1.889	76.7	76.7	76.7=(100.0×76.7)
1915	1.563	63.5	82.7	63.4=(76.7×82.7)
1916	2.507	101.8	160.4	101.7=(63.4×160.4)
1917	3.253	132.1	129.8	132.0=(101.7×129.8)

link relative has been related to the fixed base by successive multiplications. Comparing columns (2) and (4) we find very little differences in the indexes. These differences between the fixed-base indexes and the chain indexes converted to a continuous series will not always be so small. Our series is too short to illustrate how the two series may separate if continued over a long period. Errors which are negligible between two months or years cumulate in the products as the number is increased.[1] *No advantage is claimed for the chain index number in making comparisons with a fixed base* and the student is warned against its use for this purpose.

The choice of a fixed-base period. Should the base be a single year, as 1913 now in use by the Federal Bureau of Labor Statistics, or a period of years, as 1890–99 as originally used by that Bureau? Does the location of the year or period with reference to high or low prices make any difference? What should determine the selection?

The base selected should be that period with which comparisons are fair and most significant for the purpose. This principle explains the selection of the year 1913 by the Federal Bureau. The greatest interest in recent price comparisons has been concerned with the changes caused by war and post-war conditions. In selecting the period of time covered by the base, however, care should be exercised to determine a normal level from which to measure the changes. We wish to avoid an abnormal distribution of price fluctuations.

In case the price of a particular commodity is unusually high or low in the base period, its relative prices at other periods will be exaggerated.

[1] See Bulletin 284, United States Bureau of Labor Statistics, pp. 85–89, for a fuller discussion and illustration.

If the simple arithmetic method of averaging relatives is used to obtain the index very high or very low rılatives will distort the average, especially if a number of the commodities show extreme variations. This method of averaging gives relatively more influence to rapidly rising prices than to prices that advance slowly. On the contrary, it gives relatively less influence to rapidly falling prices than to slowly falling prices.[1] *These unusual prices are less likely to persist for a year than for a month and still less likely to persist over longer periods of three, five, or ten years.* Of course if the geometric mean is used to combine the relatives the influence of extreme variants is minimized. When the method of comparing aggregate values is used the dangers of averaging relatives are avoided.

Furthermore, it must be remembered that business moves in periodic waves of prosperity and depression. The prices of any short period, such as a particular month or year, may be at the peak or at the low point of the movement, or on the upward or the downward trend. If the period is made long enough to avoid both a low level and a high level by combining both high and low prices, the base will be more nearly normal than if the prices of a short period are used. It will be recalled that the financial arrangements between the Government and the railroads during the period of war control were based upon the average conditions of 1911, 1912, and 1913 and not upon a single year. *Evidently the single pre-war year,* 1913, *was not considered typical.*

Let us illustrate by Table 43 A, B, and C the differences in the indexes which may be produced by the selection of different base periods. The price data are wholesale prices of crūde oil in bulk at the wells in three of the principal oil fields of the United States.

In Table B the prices of a single year, 1913, are used as a fixed base for computing the relatives. These prices were unusually high as Table A shows. Therefore, all relatives computed from them are abnormally low and the resulting combined indexes are also low. On the other hand, in Table C the prices of three years are averaged for each field to furnish the base for relatives for that field. The resulting relatives and the combined indexes appear different being on a much higher level, because they are measured from a lower and a more typical starting point as a base.

Sometimes the mere choice of a particular base period assists the investigator to prove what he wishes to show, especially if a short period be accepted as a base and comparisons are made between several series. For example, some interested investigator desires to show that wages have kept pace

[1] A. A. Young: "Index Numbers," *Handbook of Mathematical Statistics*, pp. 182–83.

TABLE 43. COMPARISON OF INDEX NUMBERS COMPUTED ON DIFFERENT BASES
(Method of Simple Averages of Relatives)

A — AVERAGE PRICES OF CRUDE OIL AT THE WELLS (BARREL) 1911–17

FIELD	*1911*	*1912*	*1913*	*1914*	*1915*	*1916*	*1917*
Pennsylvania........	$1.301	$1.598	$2.463	$1.889	$1.563	$2.507	$3.253
Mid-Continent......	.479	.692	.951	.803	.587	1.189	1.730
California..........	.477	.454	.467	.482	.422	.590	.918

B — INDEX NUMBERS FOR THREE FIELDS COMBINED, 1913 = 100

	1913	*1914*	*1915*	*1916*	*1917*
Pennsylvania..................	100	76.7	63.5	101.8	132.1
Mid-Continent..................	100	84.4	61.7	125.0	181.9
California.....................	100	103.2	90.4	126.3	196.6
Index numbers...............	100	88.1	71.9	117.7	170.2

C — INDEX NUMBERS FOR THREE FIELDS, AVERAGE PRICES 1911–13 = 100

	Average *1911–1913*	*1914*	*1915*	*1916*	*1917*
Pennsylvania...............	100	105.7	87.5	140.3	182.0
Mid-Continent..............	100	113.6	83.0	168.2	244.7
California..................	100	103.4	90.6	126.6	197.0
Index numbers............	100	107.6	87.0	145.0	207.9

with or have exceeded the rising prices of commodities during the years preceding the peak of 1920. He may select 1917 as the fixed base for both relative prices and relative wages. Before 1917 prices had already advanced sharply while wages lagged far behind. But after 1917, when the United States had entered the World War and labor had become scarce, wages rose very rapidly while prices continued their upward trend but at a more moderate pace than wage changes. Therefore, by choosing 1917 prices and wages as 100, the index numbers for the following years, when wage changes and price changes are compared, appear far more favorable to labor than the facts warrant. A very different picture would have been shown had 1913 or 1914 been selected as a base period, *because by 1917 one of the series, compared with the other, was already on a high level.* It may be noted that some of these difficulties of selecting a proper base are avoided when the method of aggregate values is used or when the relatives are combined by the geometric mean.

Another consideration should be noted when the method of simple averaging of relatives computed on a fixed base is employed. The longer the same *fixed base* is continued the greater the spread of the variations in the price changes about their central tendency.[1] The wider dispersion which is characteristic of fixed-base relatives as time goes on leads to an increasing error in combining the relatives. Some prices fluctuate far above the average and others far below, but *the high relatives exercise much more influence upon the average than the extremely low ones.* The following example makes this clear:

	1890–99 Base = 100	*1913*
Commodity No. 1........	100	200
Commodity No. 2........	100	50
Simple Mean	100	125

Commodity No. 1 doubled in price while commodity No. 2 halved in price. If we assign equal importance to the two commodities there is no net change of prices. The index, however, becomes 125 by the method of simple averaging. Undue influence is exercised by commodities which are rapidly growing more costly. But commodities rarely gain in real importance because of a great rise in price. *The error illustrated above is in the upward direction and is likely to grow in absolute amount the farther away in time we proceed from a fixed base.*[2] The use of the geometric average to combine the relatives in the above example would yield the same index for 1913 as for 1890–99 ($\sqrt{200 \times 50} = 100$).

RELATIVE IMPORTANCE OF COMMODITIES—WEIGHTING

By weights an attempt is made to assign to each commodity an influence in the final result proportionate to its importance relative to other commodities. *Index numbers may be distinguished as simple or weighted.* In other words, if no specific system of weights is applied it is assumed that the items combined have equal importance. The examples of relatives combined by averaging presented in this chapter, have been of this type. Professor A. L. Bowley declares that "no great importance need

[1] For proof of this see Professor Wesley C. Mitchell's discussion in Bulletin 284, United States Bureau of Labor Statistics, pp. 11–23. Graphs opposite p. 14 and on pp. 19 and 20 are especially helpful.

[2] It should be noted that while the dispersion of the relatives increases as we proceed farther from the fixed base it grows at a slower rate from period to period. Therefore, the comparative error decreases, and the year to year comparisons between indexes computed by a simple average of relatives on a fixed base tend to become more accurate.

be attached to the special choice of weights," and stresses the principle that in averaging we "give all care to making the items free from bias and do not strain after exactness in weighting."

On the other hand, Professor Mitchell cautions the maker of index numbers against haphazard or careless weighting through lack of attention to the subject. A so-called simple index may turn out to be heavily weighted. For example, in the Aldrich Report of 1893, a simple wholesale price index number was published in which were included relative prices of twenty-five different kinds of pocket-knives, "giving this trifling article," as Mitchell points out, "an influence upon the result more than eight times greater than that given to wheat, corn, and coal put together." "The real problem for the maker of index numbers is whether he shall leave weighting to chance or seek to rationalize it." *In this sense there is no such thing as an unweighted index number.*

If some rational system of weighting is considered worth while, then the purpose of the investigation and the use of the index will be of fundamental importance in deciding the appropriate weights. In combining aggregate values at retail prices of a definite quantity of food, Table 41, *the average amounts consumed by a typical family were used as weights. This is the present practice of the Bureau of Labor Statistics.* Before 1914, when the Bureau used the method of averaging relatives in computing its retail food index, weights were assigned to each relative according to the average amount of money expended in a typical family budget for the particular commodity. The result was a weighted index number giving to each food commodity an influence proportionate to its relative importance in family expenditure (see Bulletin 156 for details).

In the case of some wholesale general-purpose index numbers in current use weighting is secured by a careful selection and grouping of commodities, and by including price quotations for an important raw material and its derived products, as wheat, flour, bread; iron, steel, tools; hogs (live), ham, bacon, lard. The new wholesale index of the British Board of Trade includes 150 quotations in eight groups, using market prices and the method of combining relatives by the *geometric average.*[1] The former wholesale index of the United States Bureau of Labor Statistics was weighted in this manner, and was constructed by the method of a simple average of relatives. Bradstreet's wholesale index is of the same type in reference to weighting, although it combines the various prices per pound by addition instead of by averaging. It is a sum of actual prices per pound. *Since 1914 the Federal Bureau has adopted the method*

[1] See A. W. Flux, "The Measurement of Price Changes," *Journal of the Royal Statistical Society* (March, 1921), vol. 84, part II.

of comparing aggregate values of a quantity of goods, and uses for weights the amounts of the various commodities which enter into exchange in the country's commerce.

In some cases where relatives are averaged, weights are assigned to each relative according to the *money value* of that commodity entering into exchange at the period chosen as a base. The indexes computed by this method are identical with those obtained from ratios of aggregate values. (See page 200). Quantities or values produced, consumed or exchanged have been utilized in weighting schemes for wholesale index numbers. Professor Mitchell concludes that "the quantity of the article that enters into exchange, then, irrespective of the number of turnovers, is probably the most satisfactory gauge of importance to apply in making general-purpose index numbers." The answer to the question as to whether the weights should be *sums of money or physical quantities* depends upon whether the index is being constructed by the method of averaging relatives or by the method of comparing aggregate values. Money values are appropriate weights for relatives, and physical quantities are naturally employed to arrive at aggregate values.

Bias resulting from weighting.[1] Care in the choice of weights is necessary in order to avoid bias. When money values are used to weight relatives the prices are factors in the weights (quantities × prices). If values are chosen for this purpose at the later of two periods compared these money values are likely to be directly correlated with price changes. On the other hand, the relationship will usually be inverse between the price changes and money values of the earlier of two periods compared. Therefore, while the use of the simple arithmetic average of relatives introduces a bias, as already explained, using as weights money values of the base year avoids this bias for the most part, since the bias of the weights is in the opposite direction to that of the unweighted average of relatives. Using the money values of the given year as weights for relatives reinforces the bias of the unweighted relatives.

The geometric average of relatives is sensitive to bias in the weights. This difficulty may be met by adopting as weights averages of the money values of the base year and the given year or for the series of years covered, provided the quantities are available for computing money values year by year.

The aggregative type of index numbers, as a rule, is subject to little bias in the weighting. Especially for wholesale prices, constant weights give trustworthy results in series consisting of a fairly large number of price quotations. In cost of living indexes constant weights are not so

[1] A. A. Young: *Handbook of Mathematical Statistics*, pp. 192–93.

satisfactory because the correlation between prices and quantities consumed, the weights, is usually inverse. Professor Fisher's ideal formula, presented later in this chapter, offers a way of meeting this difficulty, provided quantities consumed are available.

Changing importance of commodities. Should the weights be changed frequently or remain fixed for a long period? Weights indicate relative importance, but the relative importance of commodities is changing as time passes. Therefore, the same weights long continued lose their significance. On the other hand, if weights are changed often, two sets of variables are introduced, prices and weights, both of which enter into the resulting index. But index numbers are designed to measure changes in the price level. The chain index allows the compiler to revise the weights as often as desired since accurate comparisons are shown only for successive years or months. For fixed-base indexes it is probably best to accept a compromise between the two evils by revising the series of weights once a decade and then computing the two overlapping series of index numbers for a limited period, using the old and new weights. This procedure will show the differences in the indexes due to the system of weighting.[1]

THE USE OF AVERAGES IN CONSTRUCTING INDEX NUMBERS

It should be recalled that some compilers of index numbers do not compute averages in combining price data. For example, Gibson's index is a *sum of relative prices*. Bradstreet's index dispenses altogether with relatives and takes a *sum of actual prices per pound* of a selected list of commodities. The United States Bureau of Labor Statistics employs the method of comparing the sums of actual values of a definite quantity of goods by reducing these aggregates to relatives. However, most index numbers compiled in the past and in current use have been computed by taking an average of the relative prices of separate commodities. Which kind of average is most appropriate?

As a general rule the arithmetic average, weighted or unweighted, has been used to combine the relatives — the sum of the items divided by their number. The median has been used by some well-known authorities, especially for the purpose of avoiding the undue influence of extreme variants upon the average and because of the ease with which it may be

[1] The United States Bureau of Labor Statistics has adopted this plan. The old weights for wholesale prices are found in Bulletin 181, United States Bureau of Labor Statistics, *Wholesale Prices 1890 to 1914*, Appendix; and for retail prices of food in the *Monthly Labor Review* (November, 1918), p. 95. The present weights used for wholesale prices are in the Appendix of Bulletin 320, *Wholesale Prices 1890 to 1921;* and for retail prices of food in the *Monthly Labor Review* (March, 1921), p. 26.

determined. The economist, W. S. Jevons, in his investigations of the fall in the value of gold, used the geometric average. Mr. A. W. Flux has recently adopted this method for the new British Board of Trade index. The Harvard Committee on Economic Research uses it for some of their indexes, as do also certain Federal Bureaus and business organizations.

The chief advantages claimed for the geometric average have been illustrated already in this chapter. It is not influenced by extreme variants to the same extent as the arithmetic mean, and the index so constructed may be easily and accurately shifted to any desired base by the short method. It averages relatives without bias. There are certain objections to this method of averaging. If the index is intended for general use, this form of average is little known and, therefore, the index is more likely to be misinterpreted. *It gives equal importance to equal ratios of change,* without reference to the previous level of prices or the amounts of money represented by the changes involved. If the price of one commodity has doubled and if another has declined one half the geometric average of the two relatives leaves the level at 100. Usually we are more concerned with the money cost of goods than with the ratio of price changes. Doubling the price of one item in the family budget may prove to be much more important than halving the price of another item. Besides, the computation of the geometric mean by the use of logarithms is more laborious than the computation of the arithmetic mean or the median.

These objections gain weight since some other methods of constructing an index number have the advantages of the geometric average method without its disadvantages. The index made by the method of summing up actual prices of a definite quantity of goods and comparing these aggregates in the form of relatives can be shifted to any desired base. Medians may be used to avoid the influence of extreme variations.

Of course it is recognized that the median has limitations when the items to be averaged are few and the relatives show no marked concentration. In addition it may be emphasized that the index number so constructed cannot be shifted with accuracy to a new base by the short method. The median is not capable of algebraic treatment to the same extent as the arithmetic average. *For example, median relatives representing different groups of commodity prices cannot be combined or averaged.* This characteristic of the median becomes important in such a case as the wholesale general-purpose index of the Bureau of Labor Statistics, where the computation is made by the method of aggregate values, for several separate groups of commodities, as farm products, food, building materials, etc., and then the entire list of commodities is combined into a single index by adding together the aggregate values obtained for the groups. If medians of relatives were used for the groups, these medians could not be combined into a single index for all commodities. It would be necessary to locate the median of all the separate price relatives to secure a final index.

It may be useful to summarize at this point the advantages and shortcomings of the arithmetic average in index construction. This form of average is well understood, is clearly defined and easy to compute. Moreover, it is capable of algebraic treatment, in contrast to the median. The chief limitations of the simple arithmetic mean in combining relatives are these:

(1) It is likely to be unduly influenced by extreme variants.
(2) The index so constructed cannot be shifted with accuracy from one base to another by the short method.
(3) It introduces a serious error in the index numbers in the upward direction when a fixed base is used over a long period.

The choice of an average for use in constructing an index number from relatives requires careful consideration of the data and of the purpose of the index. Experiment with more than one method is likely to show the most useful procedure. If we wish to measure the *average ratio of change*, the geometric is the appropriate average to use. But most often our interest centers not in the average ratio of changes in prices but in *changes in money cost of goods*. For this purpose the weighted arithmetic mean or the aggregative are the logical methods to use. If the variants are extreme among the relatives to be combined, then the median may yield more representative results than the mean. The desirability of having an index which can be readily and accurately shifted to a new base must be kept in mind.

The fact that no form of average made from relatives is free from objections constitutes a strong argument for giving up the use of relative prices and for adopting wherever possible the method of weighted aggregates of actual prices. Whoever constructs an index by this method is really compelled to adopt a systematic scheme of weighting. The aggregate values of a definite quantity of goods are easy to understand and less difficult and tedious to compute because no relatives are used except in comparing two or more aggregate values as a final step in the construction of the index. The relatives made from these aggregates may be shifted to any base with entire accuracy by the short method. This is a great convenience to the investigator because he is enabled to make a direct comparison between the prices at any dates in which he is interested. Also such an index is comparable with any other index covering the same period on whatever base the latter has been computed. The meaning of these weighted aggregate values is definite and they can be manipulated mathematically. If properly weighted they are not likely to be unduly influenced by a few extreme variations in price, and yet each actual price quotation influences the result. Therefore, the method of aggregate values combines most of the advantages and few of the defects which are associated with the various methods of averaging relatives.

Furthermore, the relatives calculated from the aggregate values are identical with the arithmetic means of relative prices, when the latter are weighted by the aggregate values of the base year. *The money values used as weights in combining the relatives are secured by multiplying the physical quantities by the prices of the base year.* An illustration in Table 44 A and B, consisting of the four pork products from Table 41, makes this point clear. (See page 200.)

From the tables it is shown that the two methods yield the same result in the index for 1920 on the 1913 base (100). The weights used for the relatives in Table B are the aggregate values, column (4), Table 44 A, obtained by multiplying the actual prices of 1913 by the quantities for that year. In Table B the relatives are computed for each commodity in the usual manner, with the 1913 prices as divisors. These separate relatives for 1920 on the 1913 base are multiplied by *weights in the form of money values* of a definite quantity of goods at the prices of the base year. These products are added and the sum is divided by the total of the weights. The resulting weighted index is the same as that obtained by the method of relating aggregate values in Table A. The results must always be identical and the method of comparing weighted aggregates of actual prices is much easier to compute and simpler to present. Where the data are so heterogeneous that they cannot be added, as in some production facts, the aggregative type cannot be employed and relatives must be used. A properly weighted geometric average of these relatives gives satisfactory results.

THE IDEAL FORMULA RECOMMENDED BY PROFESSOR FISHER[1]

Some authorities, as Mr. Walsh, Professor Pigou and Professor Fisher, advocate the adoption of a more complicated procedure in the construction of general-purpose index numbers in which *the weights are changed each year.* As Professor Persons points out, such an index "measures neither the varying cost of a constant amount of goods nor the varying

[1] See Irving Fisher: *The Making of Index Numbers.*

TABLE 44. COMPARISON OF THE METHOD OF AGGREGATE VALUES AND THE METHOD OF WEIGHTED MEANS OF RELATIVES

A — INDEX FOR PORK PRODUCTS 1920 ON 1913 BASE (100)
(Method of weighted aggregates of actual prices)

COMMODITY	AVERAGE PRICE (pound) 1913	QUANTITY CONSUMED (pounds)	AGGREGATE VALUES 1913 (2)×(3)	PRICE JULY, 1920	AGGREGATE VALUES 1920 (3)×(5)
(1)	(2)	(3)	(4)	(5)	(6)
Pork chops........	$.210	114	$23.940	$.437	$49.818
Bacon.............	.270	55	14.850	.547	30.085
Ham...............	.269	55	14.795	.597	32.835
Lard..............	.158	84	13.272	.290	24.360
Aggregate values..........................			66.857		137.098
Index..................................			100		205

B — INDEX FOR PORK PRODUCTS 1920 ON 1913 BASE (100)
(Method of weighted means of relatives)

COMMODITY	AVERAGE PRICE (pound) 1913	RELATIVE 1913	PRICE JULY 1920	RELATIVE JULY 1920	WEIGHTED RELATIVES 1920 (Weights=Quantities at base year prices, column (4) Table A)
(1)	(2)	(3)	(4)	(5)	(6)
Pork chops......	$.210	100	$.437	208×	$23.940=4979.520
Bacon.............	.270	100	.547	203×	14.850=3014.550
Ham..............	.269	100	.597	222×	14.795=3284.490
Lard.............	.158	100	.290	184×	13.272=2442.048
Index........................		100			66.857)13720.608 205

amounts of goods which a dollar will buy."[1] *This type of index does not measure changes in the price level alone* and is defended on the ground that the importance of price changes depends largely upon corresponding variations in the amount of goods bought at the changing prices, a fact which is given proper influence in the index by changing the weights at each successive period compared with the chosen base.

[1] "Fisher's Formula for Index Numbers," *Review of Economic Statistics*, Harvard University (May, 1921), p. 112, note.

Professor Fisher's ideal formula for the purpose may be expressed in the form of symbols as follows:

$$\text{Price Index of Given Year or Period} = \sqrt{\frac{\Sigma p_1q_1}{\Sigma p_0q_1} \times \frac{\Sigma p_1q_0}{\Sigma p_0q_0}}$$

in which, p_1 and q_1 = price and quantity of a commodity for a given year or period; p_0 and q_0 = price and quantity of a commodity for the base year.

To use this formula it is evident that data must be available for both prices and quantities of each commodity for each year or period covered by the index numbers. *The method is that of comparing aggregate values of a definite quantity of goods.* In the first factor under the radical the weights are the constant quantities of the given year; and in the second factor the weights are the constant quantities of the base year. The two ratios computed by the method of aggregate values, the first using the quantities of the given year or period as weights, the second using the quantities of the base year or period as weights, are *averaged by the geometric method of multiplying them as factors and extracting the square root.* This is merely a device of introducing both the weights of the given year and the base year into the same index by averaging the two results in such manner as to give equal influence to the two ratios, *each ratio having been first computed by the usual method of aggregate values, already described and illustrated.*

The difficulty of securing the proper weights for each period covered by the index as well as for the base year is well nigh insuperable. Such an index, however, if it can be constructed, admirably serves as *a measure of changes in the volume of "real income"* for which Professor Pigou would use it, because the changing quantities for which incomes are expended are necessary factors. But this index on a fixed base does not measure the changes in the price level alone, as long as the two variables enter into the final result.[1]

GRAPHIC METHOD OF SHOWING PRICE CHANGES

In fixed-base index numbers changes are measured from a common base (100). A heavy line extending horizontally through the page designates this 100 level. Units of time in months or years are laid off on the horizontal scale. At each unit of time the corresponding index is plotted vertically above or below the 100 line, according to a predetermined scale. The line at 100 on the vertical scale is really the zero base line of reference, since we are interested in showing fluctuations from a fixed base. But do

[1] For comment by Professor Mitchell see Bulletin 284, United States Bureau of Labor Statistics, pp. 92 and 93.

we wish to portray absolute amounts of difference or percentage changes from period to period? Should this vertical scale be so constructed that a given distance, for example one inch, always represents on the diagram the *same absolute number of points of change* in the index, say ten points; or should the vertical scale be laid off so that the same distance always indicates the *same ratio of change from the preceding level* of the price index? The former is the *natural scale* and the latter is the *ratio or logarithmic scale.* The general explanation and illustration of these two types of diagrams are found in Chapter XVI (Figures 72, 73, 74, 75, 76, 77).

Figures 21 and 22 illustrate the two methods of presenting the same data.

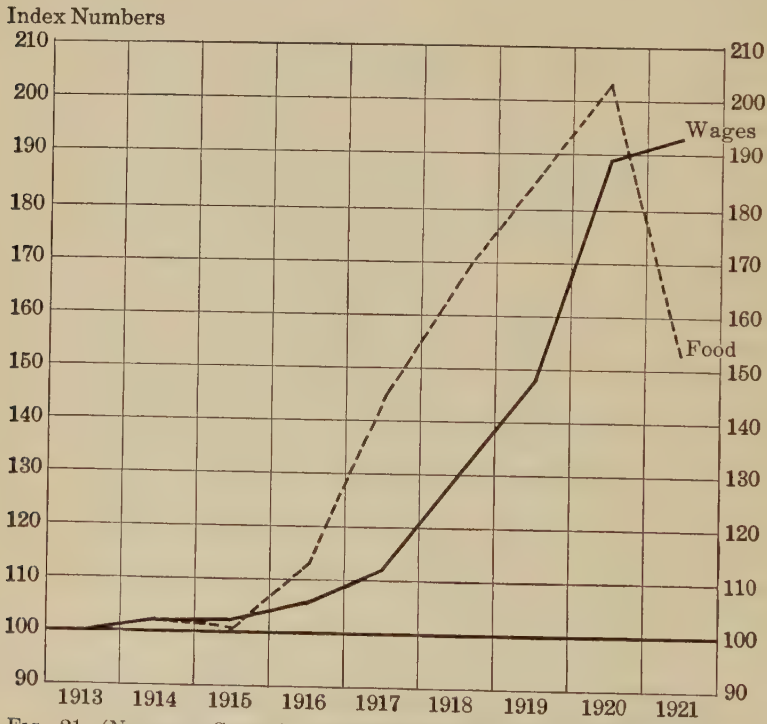


FIG. 21. (NATURAL SCALE.) COMPARISON OF CHANGES IN FULL-TIME WEEKLY WAGE RATES (CONTINUOUS LINE) WITH CHANGES IN RETAIL PRICES OF FOOD (DOTTED LINE), 1913–1921

(Data from Bulletin 315, Retail Prices, Bureau of Labor Statistics, Appendix A, Table 1, p. 179.)

In Figure 21 any point on the line diagram indicates the level of average prices of food or wages above 100 at the particular time. Equal vertical distances on the diagram represent the same absolute number of

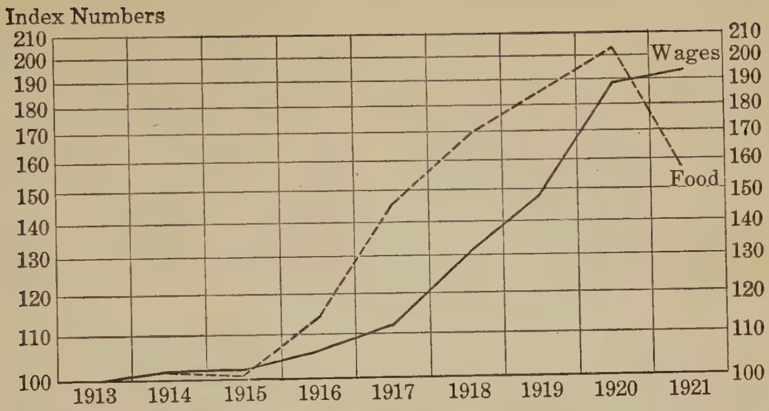


FIG. 22. (RATIO OR LOGARITHMIC VERTICAL SCALE.) COMPARISON OF CHANGES IN FULL-TIME WEEKLY WAGE RATES (CONTINUOUS LINE) WITH CHANGES IN RETAIL PRICES OF FOOD (DOTTED LINE), 1913–1921

(Same data as Fig. 21.)

points of fluctuation in the index numbers. *The slope or steepness of the line between two points on the diagram has no significance and must not be interpreted as indicating rate of increase or decrease.* Comparison of two or more lines, representing different series of index numbers on the same diagram, is likely to be misleading because the observer interprets the slopes of the lines in terms of rates of change. *Only levels or points on the lines are significant.* This is equally true in comparing different parts of the same line.

In Figure 22 equal vertical distances indicate the same percentage differences between index numbers. In this diagram *the slope or steepness of the line is of chief significance,* and indicates rate of increase or decrease. Two or more lines on the same diagram, drawn on the ratio scale, are comparable as indicating relative rates of increase or decrease. The same is true in comparing different parts of the same line.

SUMMARY

Our discussion of the principles governing the making of index numbers may be summarized. The reader is asked to remember that no attempt has been made to make the treatment exhaustive or the conclusions final. The primary purpose of this chapter has been to describe the application of statistical methods in a field where important and surprisingly accurate results are being obtained by different workers and by widely different methods. Future experimentation must settle many controversial questions.

1. As in every statistical inquiry and procedure, the purpose in view should guide in the choice of materials and methods for the construction of an index number.

2. In the field of statistical method, the index number, as a device to measure group changes, belongs in the family of averages. The index is a summary expression, convenient for representing the trend of separate price variations. Like all summary expressions *it is far from adequate to represent all the facts which are thus summarized.* Therefore, our increase of knowledge about prices and why they vary requires intensive study of the original data and experimentation, rather than reliance upon some particular form of average or aggregates or system of weighting. *It follows that it is most important at the present stage to collect accurately from representative sources and to publish in detail the actual prices of as many commodities as possible.*

3. A study of the variations of the prices of different commodities from one month or year to the next shows the tendency of these variations to concentrate fairly closely about their average. This scatter of individual price variations increases as the given period departs from the fixed base, but at a diminishing rate. Erratic variations introduce the need for caution in the choice of a particular kind of average, arithmetic, median or geometric. The difficulty of averaging relatives which vary widely and the desire for a short method of shifting the base leads to the adoption of the method of aggregate values of a definite quantity of goods or to the use of the geometric average. Index numbers tend to become less accurate as they extend over a longer period from a fixed base.

4. For the general-purpose wholesale index the best form for measuring the average change in the amount of money required to buy a definite quantity of goods seems to be the weighted aggregate of actual prices. For this purpose probably the best available weights are the quantities of the commodities bought and sold over a period of years without reference to the number of times they pass from hand to hand. In case it is desired to measure the average ratio of change in prices, without emphasis upon money cost, the geometric mean should be used. A rational system of weighting is desirable in either case.

5. Collection of price data and the selection of the representative commodities, markets and grades are excellent examples of sampling. In a general-purpose wholesale index many classes of commodities must be represented. The more commodities included in this index, the better, if care is exercised to preserve the proper relative importance between commodities and between groups. Where the field is restricted, as, for example, to retail prices of food, a smaller number of commodities is

required, but the problem of securing representative stores and grades is more difficult.

6. In contrast to the difficulty of averaging relatives representing groups of commodities having varying importance, the aggregates of actual prices, weighted by quantities produced, exchanged or consumed, may be readily added together for separate groups of commodities to form a combined index.

7. Index numbers plotted on the *natural scale* and shown in the form of a line diagram are to be interpreted in a manner different from those plotted on the *ratio scale*.[1] For the purpose of comparing the slopes of two or more lines during the same period of time, or for the purpose of comparing the relative steepness of the same line at different periods of time, the ratio diagram is the only accurate graphic representation. On the natural scale a point on the line simply means a level above or below 100, measured on the vertical scale. The slope of the line tells us nothing about the rate of change.

READINGS

Mitchell, Wesley C., Bulletin 284, Bureau of Labor Statistics, *Index Numbers of Wholesale Prices in the United States and Foreign Countries, 1921.* (Revision of Bulletin 173.) Part I, *The Making and Using of Index Numbers.*

Mills, F. C., *Statistical Methods Applied to Economics and Business*, chaps. 6 and 9.

Zizek, Franz, *Statistical Averages*, translated by Warren M. Persons, pp. 95–101.

Bowley, A. L., *Elements of Statistics*, 4th ed., part I, chap. 9.

Jerome, Harry, *Statistical Method*, chaps. 11 and 12.

Secrist, Horace, *An Introduction to Statistical Methods*, chaps. 9 and 10.

—— ——, *Readings and Problems in Statistical Methods*, chap. 8.

REFERENCES

Fisher, Irving, *The Making of Index Numbers.*

—— ——, "The Best Form of Index Number," *Quarterly Publication of the American Statistical Association*, March, 1921, pp. 533–51.

Persons, Warren M., "Fisher's Formula for Index Numbers," *Review of Economic Statistics*, Harvard University, preliminary volume III (1921), pp. 103–13. (A critical review.)

Young, A. A., "The Measurement of Changes of the General Price Level," *Quarterly Journal of Economics*, vol. 35 (1921), pp. 557–73.

—— ——, *Handbook of Mathematical Statistics* (H. L. Rietz, Editor), chap. 12.

—— ——, "Fisher's The Making of Index Numbers," *The Quarterly Journal of Economics*, vol. 37 (February, 1923), pp. 342–64.

Walsh, C. M., *The Measurement of General Exchange Value.*

Mitchell, Wesley C., *Business Cycles.*

[1] The United States Bureau of Labor Statistics, in its Bulletins and in the *Monthly Labor Review*, has adopted the ratio diagram, for the most part. Comparisons of line diagrams in these publications are of fundamental importance and ratio diagrams are less likely to mislead the reader.

Day, E. E., "An Index of the Physical Volume of Production," *Review of Economic Statistics* (September, 1920–January, 1921), vols. 2 and 3.

Flux, A. W., "The Measurement of Price Changes," *Journal of the Royal Statistical Society*, March, 1921. (Describes British Board of Trade Index Number.)

Hansen, A. H., "The Buying Power of Labor During the War," *Journal of the American Statistical Association*, March, 1922.

Douglas, Paul H., and Lamberson, Frances, "The Movement of Real Wages 1890–1918," *American Economic Review*, September, 1921. (Earlier articles in same source, by Rubinow in issue of December, 1914, and by Jones in issue of June, 1917.)

Kelley, Truman L., *Statistical Method*, chap. 13.

Davies, G. R., *Introduction to Economic Statistics*, chap. 3.

Weights used by the Bureau of Labor Statistics for the current indexes on Retail Prices, *Monthly Labor Review* (March, 1921), p. 26. (Old weights in *Monthly Labor Review*, November, 1918, p. 95.)

Weights used by the Bureau of Labor Statistics for the current indexes on Wholesale Prices, Bulletin 320, *Wholesale Prices 1890–1921*, Appendix. (Old weights in Bulletin 181, Appendix.)

Monthly Labor Review and the two series on *Retail Prices* and *Wholesale Prices*, published by the United States Bureau of Labor Statistics. (Furnish excellent current materials for the student.)

Survey of Current Business, Department of Commerce, Bureau of the Census. Published monthly. (Excellent source for data.)

[Details as to publisher and date of publication of readings and references given at the close of each chapter are found in the alphabetical list in Appendix A.]

CHAPTER XI

MEASURES OF UNRELIABILITY — PROBABLE ERROR

DESCRIPTION OF A FREQUENCY DISTRIBUTION

THE frequency table groups the original data in more compact form. Averages are summary values representing the detailed measurements. They establish central or typical values, useful for comparison and useful as norms, about which the individual values cluster with varying degrees of spread. Since the average alone does not furnish a characterization of this spread, and since very different distributions often accompany similar averages, measures of variability, both absolute and relative, are required to complete the description. By this means degrees of similarity or difference among the classified values of a frequency distribution are described in relation to the average.

In Chapter V the frequency distribution was shown in graphic form, the histogram, polygon and smoothed frequency curve. These distributions assumed different shapes according to the completeness of the data and the type of facts plotted. Data such as height approximate closely the bell-shaped symmetrical form, in which the positive and negative deviations above and below the average are about equal in number; in which all forms of the average tend to be identical in value; and in which the form of the curve on either side of the maximum ordinate is the same. Other distributions, as weight and income, were shown to depart from the symmetrical form. These were later described in the chapter on Variation as skewed toward the higher or lower values on the scale, depending upon whether the range of the positive deviations from the central value was greater than that of the negative deviations, or *vice versa.*

No single quantity can fully describe a variable. Many measurements are necessary, and before inferring anything from a series of measures, the series should be examined in the form of a graph as well as in a table. Graphic representation brings out the details of the distribution as no summary figures can do.

IDEAL FREQUENCY CURVES

Since the data from which some of the polygons in Chapter V were plotted were samples of a larger population, as height, weight and piece-rate earnings, the distributions showed irregularities due to the accidents of sampling. In the chapter on the Mode, page 133, the device of a moving average of successive frequencies was used to smooth out these irregularities in the lengths of successive ordinates or vertical distances. It was pointed out in the same connection that this smoothing could be done with more precision by mathematical methods not described in this text.

In any case the smoothed curve is not the representation of actual data but is a *generalized form — a law of arrangement* — typical of a given type of data, on the assumption that the cases have been greatly increased to include the entire population from which the sample was taken, and that the class-intervals have been indefinitely narrowed. *Science demands a means of generalization, hence the need for these ideal curves* of different geometrical forms in the interpretation of limited numbers of observations in any specific problem.

Smoothing a frequency polygon until it becomes an ideal frequency curve assumes that the original sample is *representative* of the entire number of cases from which it is chosen. The sample may be *adequate* in size and yet be *unrepresentative.* The representative character of a sample is discussed in Chapter XIV.

CAUSES OF VARIATION

Errors and their characteristics. It is essential in the present discussion to make clear the particular meaning of the term *error* in statistics, to distinguish the different kinds, and to explain the nature of errors in such manner as to suggest how to eliminate them or how to measure their amount.

In the first place *errors* must not be confused with *blunders* or *mistakes.* The latter arise from carelessness or incompetency in transcribing figures or reading values from a scale. The only remedy for these imperfections is the exercise of great care in the observation and handling of figures, and the use of checking devices.

The worker in the physical laboratory knows that there is *no such thing as an absolutely exact measurement.* He weighs an object to the nearest thousandth of a gram, but this is only an approximation. He must set his required standard of exactness and seek to measure accordingly. The relation of the diameter of a circle to its circumference is described by the constant $\pi = 3.1416$, but there is nothing absolute about the value of this decimal, since the computation may be carried to many more decimal places.

Furthermore, *measurements of the same object repeated with the greatest care do not yield the same results.* For example, let a skilled surveyor take the finest instrument available and measure as exactly as possible each of the three angles of a triangular plot of ground. When he adds the three measurements the sum will probably differ slightly from 180°. Let him measure the angles again with the same care and the difference from 180° may prove to be still greater. Since these sums differ from each other and from 180° neither can be an absolutely correct result. Yet the *most probable value* may be obtained from a number of observations *by averaging the results,* even when the true value is not known as it is in the above illustration.

Error in observation is the difference between the result of an observation and the true value of the quantity measured. We may not know the *true value* in the sense used above, but the *most probable value* may be obtained by averaging, provided the errors of the separate observations are *accidental and tend to balance each other.* Then, *error is the difference between the observed value and the most probable value. Therefore, error means a deviation, not a mistake.*

These deviations are caused by a great variety of factors which may be under the control of the investigator or which may be accidental and uncontrolled.

Kinds of errors classified. There are two distinct classes of errors:

(1) Constant, persistent, biased or cumulative errors.
(2) Accidental, variable or compensating errors.

The former type arises from causes which operate in the same manner upon successive observations. For example, the marking on the measuring instrument is wrong and every observation is in error in the same direction. Many observations do not eliminate but cumulate the total error. A very tall or a very short person may read the thermometer hanging in a fixed position at such an angle as to increase or decrease the result persistently. *The prejudices or personal equation* of the investigator may influence him to observe only the phenomena which support his views. This personal bias may be unconscious but the error is *constant* and *cumulative.* Such errors are never negligible because one does not offset another. Sometimes it is possible to eliminate constant errors by having different persons observe the same phenomenon, since the personal equations of the observers may differ in an accidental manner, as in judging a contest.

In sampling, the constant or cumulative type of error affects the *representative character* of the sample. For illustration, wage data are requested from employers by means of schedules sent by mail. Only a small percentage of the schedules are returned. It may happen that only those employers reply who are paying the higher wages or who have the best organized plants where earnings are at a high level. The picture of wage conditions is distorted. The same difficulty arises when family budgets are carefully collected by special agents, but only from housewives who keep some sort of accounts or who are especially intelligent.

Errors of the second class are of a very different kind. They are due to many variable and temporary factors operating *independently* and according to chance. Therefore, they tend to occur as frequently above as below the true measure and may be eliminated by averaging — *they tend to balance or compensate each other*. Observations are affected at the same time by optical illusions, fatigue and a variety of external conditions.[1]

It is possible to combine with accuracy observations which are subject to considerable variable or accidental errors, since, when the number of cases is large, these errors tend to balance. We can often increase the number of observations when we cannot make them less liable to this type of error. For example, in the investigation of family budgets it is useless to expect the ordinary housewife to state with great exactness the

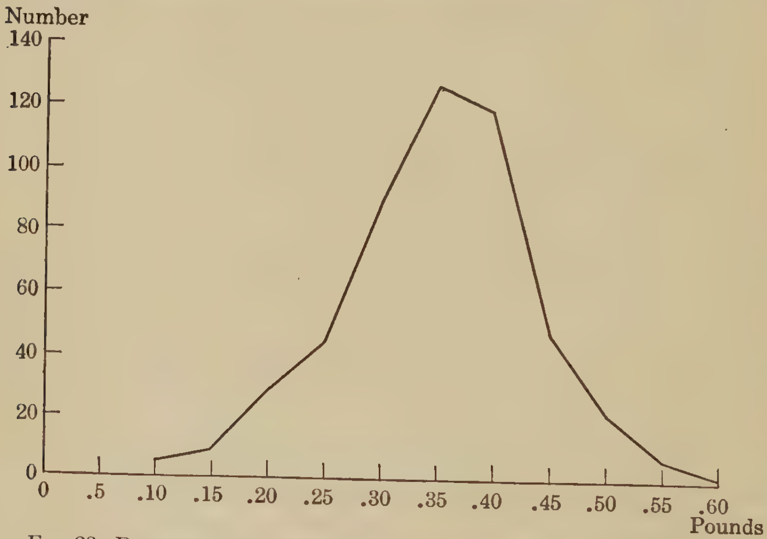


FIG. 23. DISTRIBUTION OF THE DAILY GAINS IN WEIGHT OF 498 SHEEP
(Data from Bulletin 165, Agricultural Experiment Station, University of Illinois, July, 1913, Fig. 1, opposite p. 468.)

[1] For further discussion and illustration of the subject of errors, refer to L. D. Weld, *Theory of Errors and Least Squares*, chaps. 1, 2, and 3.

quantity purchased and the amount expended for scores of items during the period of a year. However, if many budgets are collected and provided always that personal bias can be avoided, mere lack of exact knowledge on the part of the informer need not destroy the value of the results. The estimate for food made too high by one will be compensated by that of another made too low, and the average for a large number of budgets may prove to be representative. *Accidental errors are distributed about the true value or the most probable value in the bell-shaped or symmetrical form.*

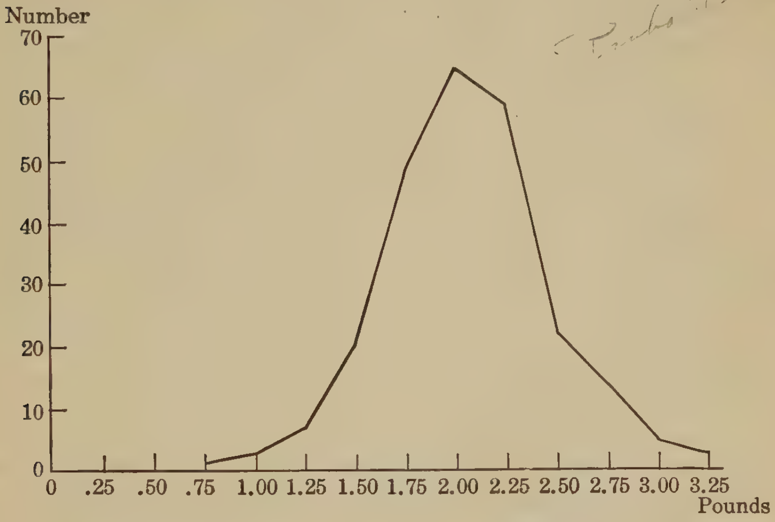


FIG. 24. DISTRIBUTION OF THE DAILY GAINS IN WEIGHT OF 241 STEERS
(Data from Bulletin 165, Agricultural Experiment Station, University of Illinois, July, 1913, Fig. 3, opposite p. 468.)

In any actual investigation some of the factors which cause variation in measurements are under the control of the observer, while others are not. For example, in measuring gains in the weight of animals fed on different diets some of the most important *constant factors which may be controlled* are the kind and quantity of the ration, its preparation, the time and method of feeding, the shelter of the animals, care in weighing, the season of the year. These may all be made fairly uniform for the given experiment. Besides these there are *variable factors* to be reckoned with which *cannot be controlled*, as temperamental differences, physiological peculiarities, feeding capacity, differential activity among the animals, weather.

Now, since the *uncontrolled* factors are numerous and act *independently* of each other, and since none are of *preponderating* influence, while each is about as likely to affect the result as any other, the frequency distribution showing the increases in weight of a number of animals approximates the bell-shaped form. Of course the gains for large lots will show a regularity of distribution which may not appear in a small number of cases. Figures 23 and 24 represent the results of actual experiments undertaken at the Illinois Agricultural Experiment Station.

PROBABILITY AND THE SYMMETRICAL BELL-SHAPED CURVE

Since the remainder of this chapter will be devoted to methods of describing and measuring accidental or variable errors, and since these errors are distributed according to the symmetrical frequency curve, it is desirable to characterize this form of distribution.

The work of Gauss and LaPlace. Working independently in physical science, they discovered that their repeated measurements of natural phenomena took the symmetrical form of distribution which they described by a mathematical equation. Gauss made repeated observations of the same phenomenon, as the diameter of a heavenly body, in order to increase the accuracy of the observation by averaging. He noted the distribution of these measurements in a symmetrical or bell-shaped form about the average or *most probable value.* Their distribution may be characterized as follows:

(1) Small deviations from the mean were more frequent than large.
(2) Positive and negative deviations were about equally frequent.
(3) Extremely large deviations did not occur.

He observed this arrangement to be in accord with the usual distributions of chance events and described the resulting frequency curve by a mathematical equation.

In the social sciences repeated measurements of the same object are not usually taken, but many records are made of the same characteristic common to many individuals, for example height, weight, age and wage. Some series approximate the symmetrical bell-shaped form, while others assume different forms. In investigations of social and economic phenomena many distributions are not symmetrical in form because, although many factors are associated with a given event, relatively few have a *predominant* influence. Moreover, these few are of unequal importance and are interrelated in such manner as to produce "skewed" distributions as the *normal or usual form for particular types of data,* as age, death-rates at different periods of life, income. It is the *unrelated* action of numerous factors which produces the bell-shaped distribution.

THE SIGNIFICANCE OF PROBABILITY

The primary function of the methods discussed and illustrated in the preceding chapters has been *description of mass data.* Numerical and graphic devices have been explained by the use of which the significant characteristics of large groups of phenomena can be briefly set forth. For the most part emphasis has been laid upon the characterization and comparison of masses of numerical data as we find them in experience.

In this chapter and the following two chapters, which deal with correlation and the time series, *the outlook is broadened.* Relationships found in experience are to be measured, and from these known observations values may be predicted which are not matters of experience. Methods are discussed which enable the student to generalize experience by extending the description of the characteristics of observed phenomena to the characteristics of similar phenomena which have not been observed. Inductive studies lead to the formulation of *general statements or laws* in which the significance of probability is important.

This point of view is in accord with Pearson's concept of causation in *The Grammar of Science.*[1] He describes causation as a stage in the routine of experience, and the concept of probability as giving expression to our belief that a certain sequence will continue to recur in the future as in our experience of the past. This belief is the basis for the prediction of future events which have not yet been observed. *In most cases our knowledge does not wait upon certainty but is described in terms of probability which may approach certainty.*

Furthermore, *the preceding chapters have stated averages and measures of variability in far too rigid a form.* These summary figures have been computed from *samples* of a larger population.[2] Other similar samples from the same population do not yield precisely the same results. It is essential for the scientific worker to describe experience in exact terms *but not with greater exactness than the facts warrant.* The concept of probability is essential in defining the degree of reliability of the various statistical measures which have been obtained from a limited number of observations — the sample. What variation may we expect *due to the chance conditions of sampling?* This is a problem of *adequacy* of the sample rather than its *representativeness.*

[1] Karl Pearson, *The Grammar of Science,* chaps. IV and V. The author wishes to acknowledge his indebtedness to Professor Pearson.

[2] The procedure of sampling is described in Chapter XIV. The reader should refer to that discussion where the problem of securing a *representative* sample is treated.

Mathematical characterization of probability. Probability may be defined as the ratio between the occurrence of an event and the group of events of which it is a part, or the ratio of the number of ways in which an event may happen to the total possible happenings, each of the ways being assumed as equally likely to occur. Since it is a purely numerical ratio it depends upon no unit of measure. For example, if an event may happen in a ways and fail in b ways, the probability of happening is $\frac{a}{a+b}$, and of failing is $\frac{b}{a+b}$. The simplest experiment is the tossing of a coin. If we toss up one penny there are only two ways in which it may fall — head or tail. Therefore, the probability of heads is one half and tails one half, which combined equal one. *Unity is the mathematical symbol of certainty.*

If an event may happen in different independent ways, *the probability of its happening is equal to the sum of the separate probabilities,* it being understood that only one of the possible events can occur. The probability of drawing a club from a deck of completely shuffled cards is $\frac{13}{52}$; of drawing either a club or a heart is $\frac{13}{52}+\frac{13}{52}=\frac{26}{52}$; of drawing either a club, a heart, a spade or a diamond is 1. The probability of drawing an ace is $\frac{4}{52}$, and of drawing an ace or a king is $\frac{4}{52}+\frac{4}{52}=\frac{8}{52}$.

Compound events. Many factors are at work to produce a given result which we measure and record quantitatively. What determines the form of the distribution of these measures has been already suggested in this chapter. Experience shows that some phenomena in society, when measured and grouped, approximate the bell-shaped distribution shown by "chance" combinations where independent events are combined.

The probability of the occurrence of a particular compound event, in accordance with chance, is equal to the product of the probabilities of the happening of the separate independent events. Continuing the card drawing illustration, if two separate packs are used and one card is drawn from each pack the probability of drawing two kings is $\frac{4}{52}\times\frac{4}{52}=\frac{1}{169}$.

To illustrate from the field of marriage statistics, suppose that in a given marriageable population 88 out of 100 possible bridegrooms are bachelors (unmarried before). Then the probability of a bridegroom being a bachelor is $\frac{88}{100}$. If 92 out of 100 possible brides are spinsters (unmarried before), the probability of a bride being a spinster is $\frac{92}{100}$. Therefore, in every 100 marriages the probability of a bachelor marrying a spinster would be $\frac{88}{100} \times \frac{92}{100} = \frac{8096}{10000} = 80.96$ per cent, *provided mere chance governs the marriages*, which assumes that *numerous independent influences* are at work to produce the given result, and that there is *no special attraction* between bachelors and spinsters. It is easy to compare the theoretical figure, 80.96 per cent, computed on the assumptions of chance, with the actual marriages of bachelors and spinsters. It will be found that the actual percentage in every hundred marriages is higher than that predicted by chance. Some special attraction must influence the event.

The more complex problem. We wish to determine the probable frequency of occurrence of compound events to which various factors operating *independently* have contributed.

In tossing a *single coin* we have seen that the probability of heads happening or failing is one half to one half. Let us introduce more coins and assume that all the coins are thrown at one time for a specified number of throws in order to determine the most probable frequency of various numbers of heads. What is the probability of no heads, one head, two heads . . . and all heads turning up in a specified number of throws?

(A) When *two coins* at a time are thrown *many times*,[1] theoretically the occurrences of heads should be

$$\left(\frac{1}{2}+\frac{1}{2}\right)^2 = \overset{0H}{\frac{1}{4}} + \overset{1H}{\frac{2}{4}} + \overset{2H}{\frac{1}{4}} = 1$$

At one of four throws no heads should turn up, at two throws one head, and at one throw two heads. The numerators of the fractions indicate the number of times heads (as shown by the symbol *H* above the fractions) may be expected to appear in the specified number of throws. In a large number of tossings of two coins no heads should appear in about one fourth of the tossings, one head in about one half of the tossings, and two heads in about one fourth of the tossings.

[1] Four throws constitute the minimum number necessary to demonstrate the possible combinations in proper proportions. When three coins are used, eight tossings are required; when four coins are used, sixteen tossings are required; and so on. Of course it should be emphasized in all the experiments here described that a large number of tossings of each number of coins is required to exhibit the proportions indicated by the fractions.

(B) When *three coins* at a time are thrown *many times,* theoretically the occurrence of heads should be

$$\begin{array}{cccccccccc} & \text{0H} & & \text{1H} & & \text{2H} & & \text{3H} & \\ \left(\frac{1}{2}+\frac{1}{2}\right)^3 = & \frac{1}{8} & + & \frac{3}{8} & + & \frac{3}{8} & + & \frac{1}{8} & = 1 \end{array}$$

In a large number of tossings of three coins, no heads should appear in about one eighth of the tossings, one head in about three eighths of the tossings, two heads in about three eighths of the tossings, and three heads in about one eighth of the tossings.

(C) When *four coins* at a time are thrown *many times,* the occurrences of heads should be

$$\begin{array}{cccccccccccc} & \text{0H} & & \text{1H} & & \text{2H} & & \text{3H} & & \text{4H} & \\ \left(\frac{1}{2}+\frac{1}{2}\right)^4 = & \frac{1}{16} & + & \frac{4}{16} & + & \frac{6}{16} & + & \frac{4}{16} & + & \frac{1}{16} & = 1 \end{array}$$

(D) When *five coins* at a time are thrown *many times,* the occurrences of heads should be

$$\begin{array}{cccccccccccccc} & \text{0H} & & \text{1H} & & \text{2H} & & \text{3H} & & \text{4H} & & \text{5H} & \\ \left(\frac{1}{2}+\frac{1}{2}\right)^5 = & \frac{1}{32} & + & \frac{5}{32} & + & \frac{10}{32} & + & \frac{10}{32} & + & \frac{5}{32} & + & \frac{1}{32} & = 1 \end{array}$$

(E) When *six coins* at a time are thrown *many times,* the occurrences of heads should be

$$\begin{array}{cccccccccccccccc} & \text{0H} & & \text{1H} & & \text{2H} & & \text{3H} & & \text{4H} & & \text{5H} & & \text{6H} & \\ \left(\frac{1}{2}+\frac{1}{2}\right)^6 = & \frac{1}{64} & + & \frac{6}{64} & + & \frac{15}{64} & + & \frac{20}{64} & + & \frac{15}{64} & + & \frac{6}{64} & + & \frac{1}{64} & = 1 \end{array}$$

(F) When *seven coins* at a time are thrown *many times,* the occurrences of heads should be

$$\begin{array}{cccccccccccccccccc} & \text{0H} & & \text{1H} & & \text{2H} & & \text{3H} & & \text{4H} & & \text{5H} & & \text{6H} & & \text{7H} & \\ \left(\frac{1}{2}+\frac{1}{2}\right)^7 = & \frac{1}{128} & + & \frac{7}{128} & + & \frac{21}{128} & + & \frac{35}{128} & + & \frac{35}{128} & + & \frac{21}{128} & + & \frac{7}{128} & + & \frac{1}{128} & = 1 \end{array}$$

In all the examples the numerators of the fractions from left to right indicate the number of times heads, as designated by the symbols above the fractions, may be expected to appear in the specified number of throws indicated by the denominators. It is understood that in each experiment a large number of tossings must be made in order to exhibit approximately these proportions.

Table 45 exhibits the results of these coin tossings.

The student should check some of these theoretical results by actual coin tossings or equivalent experiments. It is suggested that ten persons

TABLE 45. OCCURRENCES AMONG INDEPENDENT PHENOMENA

COIN TOSSINGS

Number of Coins	Number of Heads Appearing in Given Number of Tossings								Minimum Tossings
	0	1	2	3	4	5	6	7	
2	1	2	1						4
3	1	3	3	1					8
4	1	4	6	4	1				16
5	1	5	10	10	5	1			32
6	1	6	15	20	15	6	1		64
7	1	7	21	35	35	21	7	1	128
etc.									etc.

each throw seven coins at a time for one hundred and twenty-eight tossings, and average the results for each number of heads. The averages will likely approximate more closely the theoretical frequencies given in (F) than will the tossings made by one person.[1]

These experiments illustrate the nature of chance occurrences where numerous *unrelated factors* operate *independently* in producing compound events. The coins are assumed to be homogeneous in structure and, therefore, are as likely to fall on one face as the other. We toss them in such manner as to lose control of their movements when they leave our hands. A large number of circumstances can influence the movements of the coins. These factors operate *independently* and are beyond our control. Neither the structure of the coins nor the method of tossing favors one position of the coins rather than the other when they come to rest. There is no assurance that a particular coin or any number of several coins will fall heads up in any single throw, but experience shows that in the long run the coins will fall approximately according to the frequencies stated. *We expect constancy in average results.*

We are constantly making statements based upon a limited sample in terms of the most probable occurrence, which means that a generalization is made on the assumption that the sample is representative and that it is indefinitely enlarged to include cases not actually investigated.

The probability curve. Plotting the numerators in the coin experiment vertically as frequencies distributed at equal distances along the

[1] The results of this experiment are recorded in Professor H. O. Rugg's text, *Statistical Methods Applied to Education*, p. 200.

horizontal scale representing the number of heads will produce a "probability polygon." For practical purposes in checking actual distributions a continuous or smoothed curve is desired which may be so constructed as to fit the actual data of any problem (see Figure 27, page 226). This curve is known by various names — probability curve, Gaussian curve, normal curve of error, bell-shaped curve. How may it be described and how is it constructed on the basis of known data? What are its uses in practical statistical work?

Mathematical description of a curve. By "mathematical description" of any curve we mean simply giving it a name in the language of symbols as you would name a picture. The description must distinguish one type of curve from another, and it must enable us to construct a particular curve from known values and their relations. Let us illustrate this statement by simpler forms before describing the bell-shaped curve.

The straight line. The simplest form of curve is a straight line. Let us draw one in a field bounded by the coördinates OX and OY, as in Figure 25. (For explanation of rectangular coördinates, see Chapter XVI, Figure 68.)

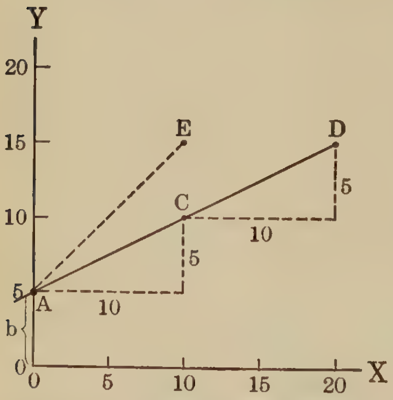


FIG. 25. GRAPHIC DESCRIPTION OF A STRAIGHT LINE

(The equation of a straight line, as AD, is $Y = mX + b$.)

The straight line AD, or any other straight line, is described by a simple equation $Y = mX + b$. This gives a *distinctive title* to the straight line as contrasted with other types of curves. What do the terms of this equation represent?

Any point on the horizontal axis is represented by X and any point on the vertical axis by Y. The values of X and Y are measured from zero origin. If we let X in the equation equal zero, then $Y = b$. The value of Y for the point located at A is 5, and when X equals zero, b has the same value, 5. We call b a *specific constant* with reference to a straight line because for a particular line, as AD, the value of b remains the same. *It is the distance from zero origin to the point where a straight line cuts the vertical axis.*

What does m represent? The point C is 5 units higher on the vertical scale than A. In passing from A to C, in the same manner as if we were climbing a hill, what horizontal distance do we traverse? As we ascend

5 units, we move to the right 10 units, or, we ascend only one half as fast as we move toward the right. *The slope of the line AC is represented by m*, and $m = \frac{5}{10}$ or $\frac{1}{2}\cdot$

The vertical distance of the point D is 5 units above C and the horizontal distance of D from C is 10 units. Again the slope of the line CD is represented by $m = \frac{5}{10}$ or $\frac{1}{2}\cdot$ The slope m of the line AD is the same throughout and, therefore, m has a constant value for this particular line, as just demonstrated. It is also a *specific constant* in the equation $Y = mX + b$.

Other straight lines can be drawn through the point A, as AE. For all lines drawn through A the value of b remains the same, but the slope of AE differs from that of AC. In the former $m = \frac{10}{10} = 1$, while for the latter $m = \frac{5}{10} = \frac{1}{2}\cdot$ However, *for any particular straight line, as AD, m has a constant value, the slope of the line between any two points located upon it.*

Suppose we wish to draw the straight line AD, knowing that it cuts the Y axis at a point 5 units from zero ($b = 5$) and that its slope is $m = \frac{1}{2}$. Of course we know also its general description by the equation $Y = mX + b$. If we can locate two points on the assumptions just stated, the line passing through these points will be the required straight line. If we assume any horizontal distance from zero as a value of X, we can compute the corresponding value for Y in the equation of the straight line. Assuming $X = 10$, $Y = \frac{1}{2}(10) + 5$, or $Y = 10$. These values of X and Y are the coördinates of the point C. To locate the point C, a line is drawn through the point at 10 on the horizontal scale at right angles to the X axis, and through the point at 10 on the vertical scale a line is drawn at right angles to the Y axis. Where these lines meet C is located.

To locate a second point, D, let $X = 20$, and we have $Y = \frac{1}{2}(20) + 5$, or $Y = 15$. By the same procedure as before, knowing values for both X (20) and Y (15), the point D can be readily located. The straight line passing through these two points is the line AD which we set out to describe. We can locate any number of points on this line AD in the same manner by assuming any values we choose for X and computing the corresponding values for Y. *The m and b for this particular line remain constant in value while the X and Y are variable values.*

Description of a circle. The reader knows how to construct a circle with a compass when the radius is known. The radius is *constant* for all points on the circumference of a particular circle. How shall we describe the circle in the language of symbols so as to distinguish it from elliptical and other types of curves?

The general equation of the circle, *its distinctive title*, is $x^2 + y^2 = a^2$, in which a is the radius, x is any distance laid off on the X axis from zero at the center, and y is any distance on the Y axis from zero, as in Figure 26.

Figure 26 presents the relations of the values in the equation $x^2 + y^2 = a^2$, when the radius, a, is 10, and, therefore, $x^2 + y^2 = 100$.

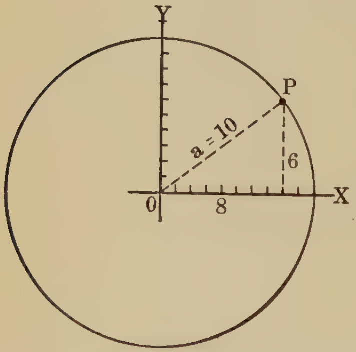


FIG. 26. GRAPHIC DESCRIPTION OF A CIRCLE
(The equation of a circle is $x^2 + y^2 = a^2$)

A sure test whether a point P is on the circumference with the center at zero and the radius 10 is to ascertain if $x^2 + y^2 = 100$. For example, if x is 8 units on the horizontal scale from zero, then $(8)^2 + y^2 = (10)^2$, and $y^2 = (10)^2 - (8)^2 = 36$, and $y = 6$. These values for x and y enable us to locate P by the same method as that described in locating C and D in Figure 25. If P is really on the circle, the equation $x^2 + y^2 = a^2$ must be satisfied, $(8)^2 + (6)^2 = (10)^2$, or $64 + 36 = 100$. *This equation is true for every point on the circumference of this circle, no matter where the point is taken, but it is not true for points inside or outside the circumference.* Any number of points on the circumference may be located by the use of this equation, with the radius (10) kept constant.

The symmetrical bell-shaped curve. It has been our purpose, in describing simple forms of curves in the language of symbols and by the use of distinctive equations, to show that there is nothing mysterious about the mathematical terms and relationships expressed in these descriptions. The student need not be perturbed about their derivation, even if he does not work out each step in the procedure.

The equation describing the bell-shaped curve is

$$y = \frac{N}{\sigma\sqrt{2\pi}} e^{\frac{-x^2}{2\sigma^2}}$$

No attempt is made to present the derivation of this equation,[1] but the meaning of the symbols should be noted carefully. The terms e and π are *constants of a more general application* than those previously illustrated and described as *specific constants* (m, b, a). The constant e is a pure number, 2.71828, which is the base of the Napierian logarithms; and π is familiar to all as the ratio of the circumference of a circle to its diameter. We understand the meaning of N and σ, representing the total frequencies and the standard deviation of the distribution. The meaning of x, deviations from the average as used heretofore, is somewhat different. It represents any distance along the horizontal scale *measured in units of* σ from the mean of the distribution as zero;[2] and y is any ordinate erected at distances x_1, x_2, x_3 .... x_n from the mean (zero).

The equation of the curve may be stated in another form

$$y = y_0 e^{\frac{-x^2}{2\sigma^2}}$$

In this equation y_0 takes the place of $\dfrac{N}{\sigma\sqrt{2\pi}}$ in the other equation and y_0 is the *maximum ordinate* of the distribution erected at the mean, or where $x =$ zero σ, forming the highest point of the bell-shaped frequency curve. It follows that the value of y_0 for any distribution is obtained from the equation

$$y_0 = \frac{N}{\sigma\sqrt{2\pi}} = \frac{N}{2.5066\,\sigma}$$

in which the terms N and σ are known for a given distribution.

When, in the equation of the curve, $x = 0$, as it will at the position of the maximum ordinate, then $y = y_0$, which is the maximum ordinate. The curve is symmetrical about the mean ($x = 0$). Since the equation $y = y_0 e^{\frac{-x^2}{2\sigma^2}}$ implies that the values of all ordinates are related to the maximum ordinate, y_0, regarded as a constant value, *the table in Appendix C has been so constructed as to give the values of successive ordinates as proportions of the maximum ordinate designated as unity.* These values of the ordinates given in Appendix C correspond to the values of x stated in amounts of σ and fractions of σ, $\left(\dfrac{x}{\sigma}\right)$, laid off on the horizontal scale from

[1] This equation describes the curve which results from $(\frac{1}{2}+\frac{1}{2})^n$ when n is made indefinitely large. See Raymond Pearl, *Medical Biometry and Statistics*, pp. 242–43.

[2] The horizontal scale is constructed by using σ as a yardstick and the mean as zero origin. The amounts of σ may be integers, as 1σ, 2σ, 3σ, or they may be fractional, as $.2\sigma$, $.4\sigma$, $.6\sigma$, $.8\sigma$.

the mean as zero. Having constructed a scale in terms of σ, the maximum ordinate from the table in Appendix C may be plotted at zero σ and the other ordinates at their respective distances from zero. The distances above and below zero for corresponding ordinates are the same. By connecting the tops of these ordinates a symmetrical bell-shaped polygon is produced. The smaller the fractional parts of σ on the horizontal scale, the larger is the number of ordinates to be plotted and the closer does the *polygon* approach to a *perfectly smooth bell-shaped curve.*[1]

Since the ordinates given in the table in Appendix C, corresponding to fractions of σ on the horizontal scale, are proportions of the maximum ordinate, unity, it is possible to multiply the maximum ordinate by any desired number, provided all the proportions of the maximum ordinate are multiplied by the same number. This procedure merely enlarges the vertical scale in plotting and does not change the form of the distribution about the maximum ordinate.

PRACTICAL USES OF THE BELL-SHAPED CURVE

The bell-shaped curve furnishes the basis for a comparison of actual with theoretical or ideal distributions. As shown earlier in this chapter, the bell-shaped curve represents a generalized experience or a law of distribution to which actual data frequently approximate. Our observations are often limited to samples from which we hope to arrive at general statements and inferences. The ideal distribution represents the situation on the supposition that the sample has been made indefinitely large. In short, it may be constructed on the hypothesis that the actual data do conform to the bell-shaped arrangement, *except as influenced by the accidental conditions of sampling* — an hypothesis which it is desired to test by comparison of the actual and the ideal distributions.

It must be remembered also that distributions may be *skewed by dominating and interrelated factors* and that in such cases departure from symmetry is not due to the limitations of the sampling procedure. The bell-shaped curve is not the appropriate form to generalize these distributions, and some other law of arrangement must be sought to describe the facts of experience. This is true of the weight and income data presented in Chapter V. *Curves other than bell-shaped have their own appropriate mathematical descriptions.*

The ideal curve superimposed upon actual data. The bell-shaped curve may be constructed to fit an actual frequency distribution for the purpose of comparison with it. The measures essential to the description

[1] For an illustration of the bell-shaped curve fitted to actual data, see Figure 27, p. 226.

of the actual and the ideal distributions are common to both, that is, the mean, the standard deviation and the maximum ordinate.

The frequency polygon of the actual distribution is plotted in the usual manner. In the ideal curve superimposed upon the actual, the assumption is made that the maximum ordinate or peak of the curve is located at the mean. The maximum ordinate, stated as unity in the table, may be magnified to the degree required, and its value for any actual distribution may be calculated from a knowledge of N and σ, by substituting these values which have been computed for the actual data in the equation $y_0 = \frac{N}{2.5066\,\sigma}$. When the maximum ordinate or frequency is known, the other frequencies may be easily computed from the table by using the fractions corresponding to the subdivisions of σ on the horizontal scale. This procedure gives the ordinates or frequencies needed for plotting the ideal curve. The reason for laying off the horizontal scale in fractions of σ in constructing the ideal frequency curve will appear more clearly from the example.

An example of fitting the bell-shaped curve. The data used for illustration are heights of Japanese and American army recruits classified in intervals of one inch. Tables 46 and 47 show both the actual and ideal frequency distributions.

The measures needed for constructing an ideal bell-shaped curve over each of these actual distributions are N, the mean and σ, which have been computed for the actual distributions by the usual short method. To obtain the *maximum ordinate or frequency* for each of the ideal distributions it is only necessary to substitute these values in the formula $y_0 = \frac{N}{2.5066\,\sigma}$[1] as follows:

$$\text{(1) Japanese } y_0 = \frac{10{,}000}{2.5066 \text{ times } 2.25} = 1773 \text{ maximum ordinate}$$

$$\text{(2) American } y_0 = \frac{10{,}000}{2.5066 \text{ times } 2.20} = 1813 \text{ maximum ordinate}$$

It is necessary to decide the values of x on the horizontal scale in terms of σ at which successive ordinates or frequencies will be plotted to secure the ideal curve. We shall use intervals of $.2\,\sigma$ for marking off the horizontal scale, beginning at the mean of the distribution as zero σ, and establishing identical points above and below the mean.

[1] It should be noted that σ in this formula must be stated in class-intervals rather than in the units of value of the particular problem, in order to obtain for y_0 a number of observations per interval comparable with the table of actual frequencies. (See G. U. Yule, *An Introduction to the Theory of Statistics*, 6th ed., 1922, p. 308.)

TABLE 46. HEIGHTS OF JAPANESE AND AMERICAN SOLDIERS

ACTUAL DISTRIBUTIONS

CLASS LIMITS (inches) (1)	MID-VALUE m (2)	JAPANESE f (3)	AMERICAN f (4)	OVERLAPPING[a] FREQUENCIES IN THE TWO DISTRIBUTIONS (5)
55.5 and under 56.5	56	47		
56.5 “ “ 57.5	57	125		
57.5 “ “ 58.5	58	316		
58.5 “ “ 59.5	59	640		
59.5 “ “ 60.5	60	1065		
60.5 “ “ 61.5	61	1486		
61.5 “ “ 62.5	62	1730	38	38
62.5 “ “ 63.5	63	1698	192	192
63.5 “ “ 64.5	64	1328	538	538
64.5 “ “ 65.5	65	839	1055	839
65.5 “ “ 66.5	66	442	1557	442
66.5 “ “ 67.5	67	208	1822	208
67.5 “ “ 68.5	68	64	1695	64
68.5 “ “ 69.5	69	12	1294	12
69.5 “ “ 70.5	70		868	
70.5 “ “ 71.5	71		510	
71.5 “ “ 72.5	72		263	
72.5 “ “ 73.5	73		114	
73.5 “ “ 74.5	74		42	
74.5 “ “ 75.5	75		12	
		10,000	10,000	2333

Japanese
N = 10,000
Mean = 62.24 inches
σ = 2.25 inches

American
N = 10,000
Mean = 67.51 inches
σ = 2.20 inches

[a] These frequency distributions overlap at heights 61.5 to 69.5 inches. This information is important in judging the significance of the means. The overlapping frequencies are given in column (5). The data are taken from F. L. Hoffman, *Army Anthropometry and Medical Rejection Statistics*, 1918, p. 33.

Now, having the maximum ordinate, y_0, for each distribution, and having established the successive horizontal distances from the mean in terms of σ, it is easy to compute the other ordinates or frequencies which are to be plotted at these points by multiplying the constant value of the maximum ordinate by the fractions in Appendix C which correspond to the distances from the mean as zero. The frequencies are given in Table 47.

The actual distribution in Table 46, column (3), is plotted in Figure 27 according to the usual method. The heights in inches are shown on the horizontal scale and the frequencies on the vertical scale as ordinates. A dotted line connecting the tops of the ordinates completes the polygon. Superimposed upon the same diagram for comparison are plotted the

TABLE 47. HEIGHTS OF JAPANESE AND AMERICAN SOLDIERS

FREQUENCIES OR ORDINATES FOR IDEAL CURVES

x/σ DISTANCE PLUS OR MINUS FROM MEAN (Mean = 0) (1)	JAPANESE COMPUTED FREQUENCIES OR ORDINATES $y_0 = 1773$ (2)	AMERICAN COMPUTED FREQUENCIES OR ORDINATES $y_0 = 1813$ (3)
0 σ = Mean	1773×1.000[a] = 1773	1813×1.000[a] = 1813
.2 σ (+ or −)	" × .980 = 1738	" × .980 = 1777
.4 σ "	" × .923 = 1636	" × .923 = 1673
.6 σ "	" × .835 = 1480	" × .835 = 1514
.8 σ "	" × .726 = 1287	" × .726 = 1316
1.0 σ "	" × .607 = 1076	" × .607 = 1100
1.2 σ "	" × .487 = 863	" × .487 = 883
1.4 σ "	" × .375 = 665	" × .375 = 680
1.6 σ "	" × .278 = 493	" × .278 = 504
1.8 σ "	" × .198 = 351	" × .198 = 359
2.0 σ "	" × .135 = 239	" × .135 = 245
2.2 σ "	" × .089 = 158	" × .089 = 161
2.4 σ "	" × .056 = 99	" × .056 = 102
2.6 σ "	" × .034 = 60	" × .034 = 62
2.8 σ "	" × .020 = 35	" × .020 = 36
3.0 σ "	" × .011 = 20	" × .011 = 20

[a] To simplify the computations only three decimals of the fractions in Appendix C are used. It should be noted that the ideal frequency curve extends beyond 3.0 σ plus and minus, but for the sake of simplicity the computations have not been made.

frequencies computed for the ideal distribution, Table 47, column (2). The points on the horizontal scale for plotting these frequencies are established by measuring unit distances, *plus and minus*, from the mean 62.24 inches (for the Japanese soldiers) as zero, *where the maximum ordinate is erected.* As shown in Table 47, column (1), these intervals are .2 σ. The value of σ is 2.25 inches. Therefore, in terms of the scale already used for plotting the actual distribution, the point at which the first ordinate above and below the mean is to be erected is determined by measuring the distance .2 times 2.25 inches, or .45 inches plus and minus from 62.24 inches. The position of the second ordinate is .4 times 2.25 inches, or .90 inches plus and minus from 62.24 inches, and so on along the scale to 3 σ, or 3 times 2.25 inches plus and minus from 62.24 inches. Having established the positions above and below the mean of the successive ordinates, we plot the maximum frequency, 1773, at 62.24 inches, and the other ideal frequencies above and below it, according to the same vertical scale used for the actual distribution. The tops of the ordinates are connected by a smooth continuous line forming the *bell-shaped symmetrical curve.* If smaller fractions of σ had been used as intervals, the number of ordinates would have been increased, and the smooth curve connecting the tops of the ordinates could have been drawn more easily.

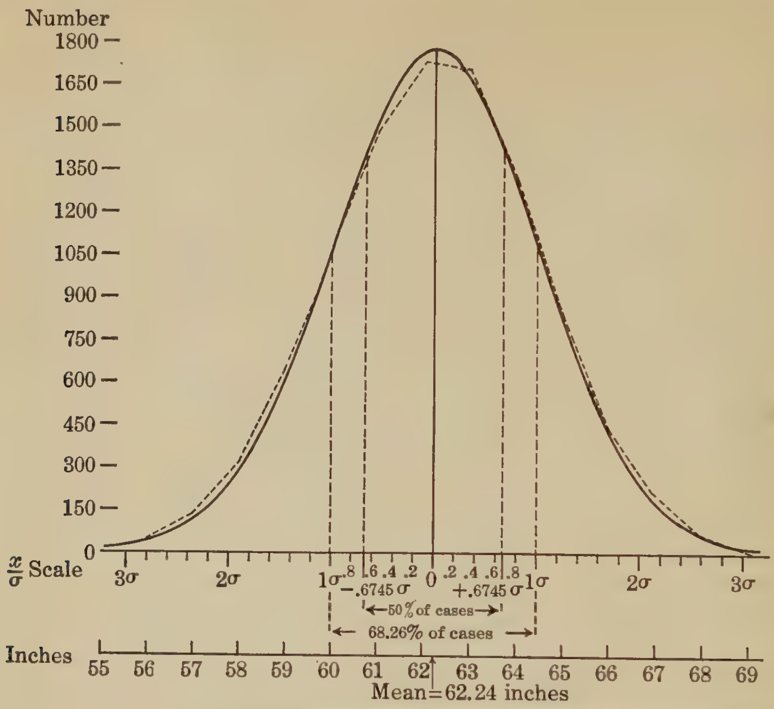


FIG. 27. ACTUAL AND IDEAL FREQUENCY DISTRIBUTIONS OF THE HEIGHTS OF JAPANESE SOLDIERS

(Data from Table 46, column (3), and from Table 47, column (2), and σ = 2.25 inches.)

The ideal curve may be extended to higher values of σ plus and minus. Between the ordinates at 2 σ plus and minus are included 95.46 per cent of the cases; between the ordinates at 3 σ plus and minus are 99.73 per cent of the cases; and between the ordinates at 4 σ plus and minus are 99.99 per cent of the cases.

Comparison of the polygon with the curve shows a very close fit and suggests that the differences between the actual and the theoretical frequencies, in the case of the heights of Japanese soldiers, *is not a real difference but is due to accidental conditions of sampling. In other words, height frequencies very probably conform to the law of distribution represented by the bell-shaped probability curve.*

The actual and ideal frequencies for American soldiers are plotted in Figure 28 from Table 46, column (4), and Table 47, column (3), in the same manner as explained for the Japanese. The intervals for the ideal curve are .2 σ. The value of σ for the American distribution is 2.20 inches and the maximum ordinate is 1813, erected at the mean, 67.51 inches.

Comparison of the Japanese and American polygons and curves shows

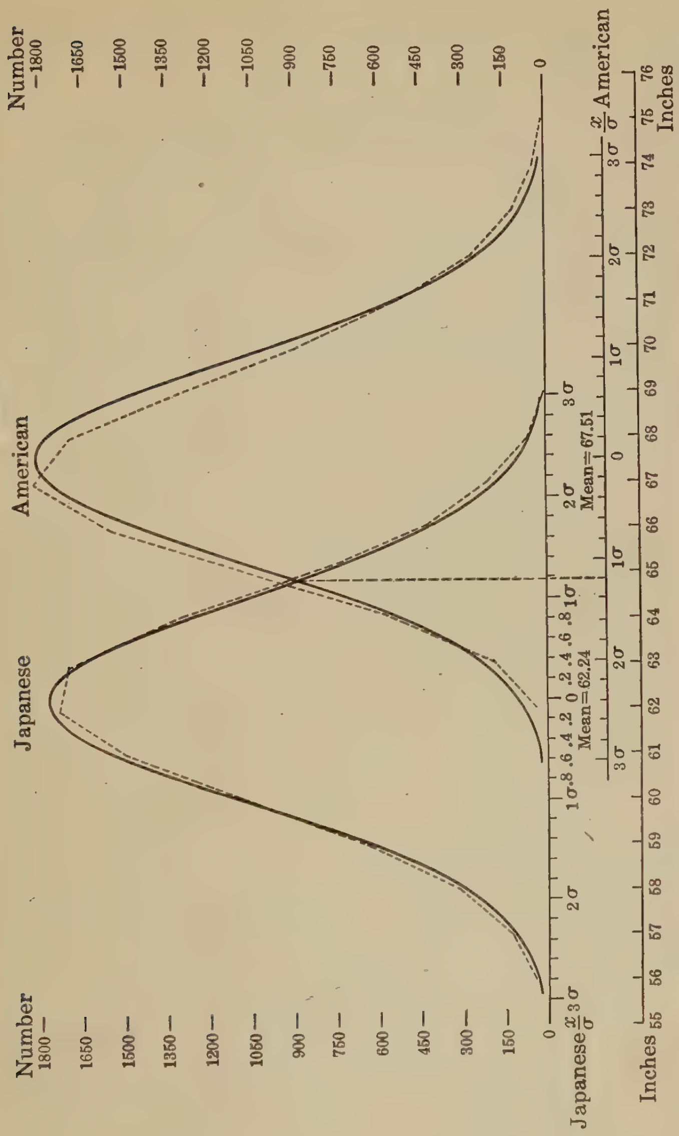


FIG. 28. ACTUAL AND IDEAL FREQUENCY DISTRIBUTIONS OF THE HEIGHTS OF JAPANESE AND AMERICAN SOLDIERS

(Original data represented by dotted polygons, and ideal data by smooth continuous lines forming bell-shaped curves. Data from Tables 46 and 47.)

that the actual distribution for the latter does not conform quite so closely to the bell-shaped form, but the difference is not so great as to indicate a different law of distribution of heights. As a further test of the symmetrical character of the actual distributions, the median and *mode* of each should be computed for comparison with the *mean.* They will be found to be almost identical in value. *In a perfectly symmetrical distribution they are identical.*

It is suggested that the student carry out the calculations necessary to construct the ideal polygon for the weight of freshmen in Table 31, page 161, on the assumption of a bell-shaped arrangement of the distribution about the mean.[1]

Comparison of the actual and ideal polygons shows that the *weight distribution does not conform to the bell-shaped curve.* The differences are not such as can be explained by accidental conditions of sampling. This indicates that weight distributions, in contrast to height shown in Figure 28, follow a *different law of arrangement, regardless of the number of cases included in the sample. Since this is true the bell-shaped curve cannot be used to generalize a limited number of weight observations.*

A question arises in connection with the comparison of actual and ideal distributions. Can we determine with precision the *closeness of fit* and decide whether the differences found at any point between the actual and the theoretical frequencies represent *real differences* in the form of the distributions, or are merely such as to be expected in dealing with a limited sample from a larger group?[2] It also raises the fundamental question of the variations which may be expected in successive similar samples *due to accidental conditions, in connection with which the size of the sample is important.* The *adequacy of the sample* must be distinguished from its *representative character.* With this in view an explanation of the *distribution of errors* will be given in the following pages of this chapter.

DESCRIPTION OF THE DISTRIBUTION OF ERRORS[3]

In Chapter IX a measure of variation was defined as a *distance on the scale within which a known proportion of the frequencies fall.* It was stated, furthermore, that within once the standard deviation above and below the mean are located about two thirds of the total items, that 6 σ should include at least 99 per cent of all the cases, and that the value of the semi-interquartile range is about two thirds of the standard deviation. The reader is asked to go over these points in review. The statements are based upon the characteristics of the symmetrical bell-shaped

[1] In calculating the maximum frequency by the formula $y_0 = \frac{N}{2.5066\,\sigma}$ it is necessary to express σ in intervals, not in pounds.

[2] For test of *goodness of fit* refer to G. U. Yule, *Introduction to the Theory of Statistics*, 6th ed., 1922, pp. 308–09, and Supplement III; and *Handbook of Mathematical Statistics* (H. L. Rietz, Editor), pp. 78–81.

[3] Errors are used in the sense of deviations.

curve with which we are now concerned. *It is possible in this chapter to make clearer the significance of the measures of variation by the use of this curve.*

Division of the area under the probability curve. Just as the maximum ordinate in the generalized curve is assumed to be unity, so the area under the curve may be assumed to be unity. The proportion of this area between the maximum ordinate and any other ordinate erected at distances from the mean expressed in terms of σ can be calculated by the *methods of the integral calculus.* This has been done in Appendix D. The figures of this table state the proportion of the total area found between the *maximum ordinate* and *any other ordinate* erected at a distance from the mean expressed in terms of σ. The areas between the corresponding ordinates above and below the mean are identical in the symmetrical bell-shaped curve.

To illustrate how the table may be used, the area of the ideal curve for the Japanese soldiers, Figure 27, may be divided. Under this curve are included 10,000 cases. To represent the area between the maximum ordinate and the ordinate erected at .2 σ, or .45 inches, above the mean the table gives the figure *0793*, which means $\frac{793}{10,000}$ of the total area under the curve, or 7.93 per cent of the total cases are found in this area bounded by the maximum ordinate and the ordinate .2 σ distant. The area bounded by the maximum ordinate and the ordinate .4 σ, or. 90 inches, above the mean is stated as *1554* in the table, meaning that $\frac{1554}{10,000}$ or 15.54 per cent of the cases are located in this area of the diagram. The area extending from the mean to the ordinate erected at 1 σ, or 2.25 inches, above the mean is given as *3413*, indicating that $\frac{3413}{10,000}$ or 34.13 per cent of the cases are located between the mean and 1 σ.

The same proportions of the cases are located within similar areas above and below the mean. Therefore, 2 times 34.13 per cent, or *68.26 per cent of the cases are located within the standard deviation measured above and below the mean.* This is the basis for the statement in Chapter IX that about two thirds of the cases are located within this distance from the mean, *plus and minus.* Likewise for 3 σ the table gives *4986.5*, which indicates that 49.86$\frac{1}{2}$ per cent of the cases are located within the area extending from the mean to the ordinate erected at 3 σ above. Therefore, 2 times 49.86$\frac{1}{2}$, or *99.73 per cent of the cases fall within a range of 6 times the standard deviation.*

What is the relation between the value of the semi-interquartile range within which 25 per cent of the cases fall and the value of the standard deviation? In other words, what fraction of σ measured from the maximum ordinate will include $\frac{2500}{10,000}$ or 25 per cent of the area of the curve? To answer the question it is necessary to find 2500 in the table, or as near to it as possible, and then note what fraction of σ includes the 2500. Follow down the second column of the table until you find the number just less than 2500. This is *2257* which is opposite .6 σ in the first column. Now follow across the horizontal row in which 2257 is located until you reach the figure just less than 2500, which is *2486*. This figure is in the column headed .07 σ. The next figure, *2518*, in column .08 σ is too large. Therefore, the figure 2500 which we wish to locate lies between the columns .07 σ and .08 σ and interpolation is necessary. The difference between 2486 and 2518 is 32 points which represents the change from .07 σ to .08 σ, or .01 σ change. The difference between 2486 and 2500 is 14 points, and this difference is $\frac{14}{32}$ of .01 σ, or .0044 σ, to be added to .07 σ, which equals .0744 σ. This fraction is combined with .6 σ, found in the first column opposite the row in which 2500 has been located, making .6744 σ. This fraction is usually given as .6745 σ, obtained by greater exactness in the decimals.

The area included between the mean and the ordinate at .6745 σ is 25 per cent of the entire area under the curve, and is exactly the same as that included within the *semi-interquartile range*. Within this distance of the mean, plus and minus, 50 per cent of the cases are found, and the chances are exactly even that any value taken at random will be located within this range of value or outside of it. The probability is one half to one half, as in the case of the penny tossing. From the above explanation the reason is clear for the statement in Chapter IX that the semi-interquartile range is about two thirds of the value of the standard deviation.

Since the table in Appendix D is constructed in terms of proportions it is applicable to any bell-shaped curve. It is possible by its use to determine directly the proportion of the values included under any such curve between the ordinate erected at the mean and any other ordinate. It is also possible to obtain the proportion of the measures located between any two ordinates on the same side of the mean, by a simple process of subtraction. For example, between the mean and .4 σ 15.54 per cent of the items are located, and between the mean and .2 σ are found 7.93 per cent. Therefore, between the ordinate erected at .2 σ and the ordinate at .4 σ the proportion is 15.54 – 7.93, or 7.61 per cent. When the ordinates between which we wish to know the proportion of cases are located on opposite sides of the mean, the process is one of addition.

The overlapping of frequency curves. A majority of Japanese soldiers were at least as tall as the smallest American. Likewise, a majority of the Americans did not exceed the tallest Japanese. The actual overlapping of the distributions is important in judging the significance of the mean height, which is more than 5 inches greater for the Americans. The frequencies duplicated in both actual distributions are tabulated in column (5) of Table 46 by noting the smaller of the two frequencies which appear at a given value in the table. The total frequencies duplicated amount to 2333 items, or 23.33 per cent of the area of each histogram.

In the ideal curves the overlapping may be computed from Appendix D in the same manner as explained in the preceding section.

Japanese curve (Figure 28).
Area of curve above the mean = 50 per cent of total area. Intersection with American curve at 1.18 σ above the mean. Mean as zero $\sigma + 1.18\ \sigma = 38.10$ per cent of total area. Fifty per cent − 38.10 per cent = 11.90 per cent duplicated.

American curve (Figure 28).
Area of curve below the mean = 50 per cent of total area. Intersection with Japanese curve at 1.2 σ below the mean. Mean as zero $\sigma - 1.2\ \sigma = 38.49$ per cent of total area. Fifty per cent − 38.49 per cent = 11.51 per cent duplicated.

Therefore, the *total duplication* in the theoretical bell-shaped curves is 11.90 per cent + 11.51 per cent = 23.41 per cent, as compared with 23.33 per cent in the actual distributions. This is a difference of only 8 cases in 10,000.

UNRELIABILITY AS MEASURED BY THE BELL-SHAPED SYMMETRICAL CURVE

The probability curve is employed to describe the probable reliability of statistical measures, as averages, measures of variation, and the measures of relationship to be discussed in the following chapter. This is the mathematical basis of sampling. It is concerned with the adequacy of the sample and with the variations due to accidental conditions. The procedure in sampling which seeks to avoid constant errors by random selection and by the elimination of bias in order to obtain representative results is described in Chapter XIV.

For example, an average is computed from a limited number of observations. The result is not likely to be identical with that which would be obtained by combining all the possible measures of the characteristic in question. Assuming that the cases have been chosen in such manner as to be representative, another sample of the same size and chosen in a similar manner will be likely to yield a different average. The same is true of other statistical measures, as the standard deviation. In other words, *these statistical measures are themselves variables.*

It is not in the interest of scientific exactness to describe distributions with greater precision than the facts warrant. It is essential, therefore, to know how the measures obtained from limited samples are distributed. If successive samples of the same kind of data are chosen at random there is no constant factor causing averages or other computed measures to fall above or below the values which would be obtained if the entire population were included. Therefore, the distribution of these measures will follow the laws of probability already described — the bell-shaped curve which is sometimes called the *normal curve of error.* The amount of *unreliability* may be stated in terms of this *probable error.*

Absolute certainty concerning statistical measures obtained from samples is impossible. When we have calculated them from a representative sample we must also define the limits within which we expect variations about these representative values to occur on account of conditions which are not within our control. The same investigator or others will examine similar samples of the same population and will compute statistical measures different from those obtained before. Are these differences significant of actual differences in the phenomena investigated or do they fall within the limits within which accidental variations may be expected to occur? If the differences are greater than can be accounted for by accidental errors of sampling, the observer has discovered something which should be explained. This is why we attempt to define a range of variation about the measure obtained from a sample within which accidental variations may be expected to occur according to specific probabilities.

An experiment with the heights of 1000 freshmen. The heights were written each upon a separate metal-rimmed disc. These were thoroughly shuffled and a sample of 100 was taken *at random.* The mean height was computed by simply adding the 100 values ungrouped and dividing by 100. The sample was restored and the entire lot was shuffled again, after which a second sample of 100 was drawn and the mean computed in the same manner. This procedure was repeated until 100 samples had been taken, each from the entire 1000 cases, and each sample of the same size. In this manner 100 mean heights were obtained from as many separate samples taken from the same general population (Tables 48 and 49).

TABLE 48. MEANS OF 100 SAMPLES — AN ARRAY OF MEAN HEIGHTS

MEAN HEIGHT (inches)				
66.92	67.30	67.43	67.55	67.70
67.03	67.30	67.45	67.56	67.70
67.07	67.30	67.45	67.56	67.70
67.07	67.30	67.45	67.56	67.70
67.08	67.32	67.46	67.57	67.72
67.08	67.32	67.46	67.58	67.72
67.10	67.34	67.48	67.59	67.74
67.13	67.35	67.50	67.62	67.75
67.15	67.35	67.50	67.63	67.77
67.18	67.36	67.50	67.63	67.77
67.18	67.37	67.52	67.63	67.78
67.21	67.37	67.52	67.64	67.79
67.22	67.37	67.53	67.66	67.80
67.22	67.37	67.53	67.67	67.83
67.23	67.37	67.53	67.67	67.88
67.25	67.37	67.53	67.67	67.92
67.26	67.39	67.54	67.68	67.96
67.27	67.40	67.54	67.69	68.02
67.29	67.42	67.55	67.70	68.16
67.29	67.43	67.55	67.70	68.23

We have treated these 100 means by the same methods used for the individual measurements within any single sample. The mean of means and the standard deviation of means have been computed from the ungrouped data in Table 48. An inspection of Tables 48 and 49 shows that the means from successive samples vary within narrow limits, that these variations group about the mean of the means, 67.50 inches, in a fairly symmetrical distribution. The irregularities in the frequencies would probably be smoothed out by more samples taken in the same manner.

How closely does this distribution of errors in the mean conform to the bell-shaped curve? The manner of selecting the samples at random emphasizes the operation of accidental, uncontrolled factors in sampling as causing the variations. In the ideal probability curve we have shown that 68.26 per cent of the items are included within a range of once the standard deviation measured plus and minus from the mean. Let us test the distribution of the ungrouped 100 means by this standard. The σ is .24 inches. Therefore, within the range 67.50 ± .24 inches (67.26 inches through 67.74 inches) should fall 68.26 per cent of the sample

Table 49. Frequency Distribution of the Means of 100 Samples

Class-Interval (1/10 inch)	Distribution of Means * = one mean	Frequency
66.90–66.99	*	1
67.00–67.09	*****	5
67.10–67.19	*****	5
67.20–67.29	*********	9
67.30–67.39	*****************	17
67.40–67.49	**********	10
67.50–67.59	********************	20
67.60–67.69	***********	11
67.70–67.79	**************	14
67.80–67.89	***	3
67.90–67.99	**	2
68.00–68.09	*	1
68.10–68.19	*	1
68.20–68.29	*	1
		100

Mean of 100 averages ungrouped = 67.50 inches
σ " " " = .24 inches
P.E. " " " = .6745 σ = .16 inches

means. Tabulating the actual means which fall within this range we find 71 out of the 100, or 71 per cent included. It will be noted that this includes one sample at exactly 67.26 and one at 67.74.

In describing the bell-shaped symmetrical curve we also showed that .6745 σ measured plus and minus from the mean should include 50 per cent of the items (Figure 27). The probability of any mean from a random sample falling within this range of value or outside of it is one half to one half. Therefore, .6745 σ measures the probable error (P.E.) of the mean. Let us test our 100 means by this standard.

$$.6745\,\sigma = .6745 \text{ times } .24 \text{ inches} = .16 \text{ inches, P.E.}$$

If the 100 sample means conform to a perfectly bell-shaped curve of distribution we should find 50 per cent of them included within the range 67.50 ± .16 inches, or between 67.34 and 67.66 inches. Tabulating the actual means within this range of values we find 47 out of 100, or 47 per cent included.

It should be observed also that all the sample means except one are included within the range 3 times .24 inches, or .72 inches plus and minus from 67.50 inches, or between 66.78 and 68.22 inches (6 σ). This also conforms to the characteristics of the bell-shaped probability curve.

The mean of the entire 1000 heights from which these samples were taken is 67.57 inches, which does not exactly coincide with the average of the 100 sample means, 67.50 inches. If many more or somewhat larger samples had been taken, these values would practically coincide, since the variation in the means is due to accidental conditions of sampling. Usually when the method of sampling is employed there is no way of knowing the value of the mean of the entire population. We only know, as indicated by the height samples, that *accidental deviations from the mean of the means due to sampling are generally distributed symmetrically about the mean of the means as a mode or point of concentration.*[1] We have just shown that the *mean of the means* (67.50 inches) approximates very closely to the *mean of the entire 1000 cases* (67.57 inches). If the experiment with the height data were continued this difference would grow smaller and smaller. *It follows that the mean of any sample chosen at random from a larger population is more likely to be located at the mean of the entire population than at any other specified value.*

The unreliability of an average. From this experiment it is clear that an average obtained from a sample is not an unvarying quantity but is subject to fluctuations over an interval on the scale. This must be considered its *unreliability,* which is described by the probable distribution of the means of a very large number of samples about the mean of all the samples. This is called the *error* of the mean. If the particular measures in each sample are chosen at random, as in the illustration, the distribution of the means obtained from successive similar samples conforms closely to the *probability curve.* The laws of probability account for the differences in the constitution of any limited group chosen at random from a larger population, and in turn produce the differences between the means and other measures obtained from samples. Therefore, in determining the unreliability of an average computed from samples, it is essential to describe the form of distribution of a large number of possible means from similar samples and to measure in terms of the *standard deviation* and *probable error* the expected variability of these sample means from their central value, the mean of the means, as in the illustration of heights.

Use of the probability curve to describe unreliability. The variability of the means computed from the 100 height samples shows a standard deviation of .24 inches measured from the mean of the means (67.50

[1] It should be emphasized that averages and other statistical constants describing representative samples of *skewed* distributions, as weight and income, will themselves generally be distributed according to the bell-shaped symmetrical form. This has been verified empirically and is logical, since the variability here described is due to the accidental, uncontrolled conditions of sampling.

inches). In a large number of samples we may assume that the mean of the means approximates very closely the mean of the entire population, as demonstrated by the height samples. It is possible to state what proportion of any number of sample means may be expected to fall within the range of 1 σ (.24 inches in the illustration), plus and minus from the central value (67.50 inches). Referring to the ideal curve for the Japanese soldiers (Figure 27) and the explanation of the division of the probability curve given on page 229, let us describe the distribution of mean heights according to the bell-shaped form. We may expect 68.26 per cent of the sample means to fall within 1 σ measured plus and minus from the mean of the means. If 10,000 such samples of height were taken, 6826 of the resulting means would probably fall within this distance of 1 σ plus and minus from the mean of all the samples. Therefore *the chances of any one of these means from a sample chosen at random falling within this same range of variation, as compared with its chances of falling outside this range, would be 6826 to 3174, or about 2 to 1.*

Likewise we may expect 95.46 per cent of a large number of sample means to fall within 2 σ (2 times .24 inches = .48 inches in the illustration of 100 samples of height), measured plus and minus from the mean of all the samples. In this case 9546 sample means out of 10,000 would probably fall within this range of variation. Consequently the chances of any random sample mean falling within or without this range of 2 σ are 9546 to 454, or about 21 to 1. Furthermore, 3 σ (3 times .24 inches = .72 inches) will include 99.73 per cent of the sample means, and the chances of any sample mean falling within or outside this range of variation plus and minus from the mean of all the samples are 9973 to 27, or about 369 to 1.

We have described the unreliability of average height on the theory that the variations of a very large number of sample means, due to accidental conditions of sampling, are distributed about the mean of all the samples [1] as a mode in the form of the probability curve. For this purpose the amount of the probable variation must be given in terms of σ or *P.E.*, the latter being .6745 σ.

The necessity for a shorter method. In practice we cannot take a large number of similar samples of the same size from the population under investigation, as was done in the illustration of heights. Yet we desire to describe the unreliability of various statistical measures computed from the given sample selected for investigation. When we use

[1] It should be remembered that the mean of all the samples approximates closely the mean of the entire population from which the samples are taken, as shown in the height samples.

these measures it is necessary to make allowance for errors which are not due to constant factors but which are related to the size of the sample and the accidental conditions of sampling. How can we use the measures obtained from a single sample to describe and to measure the probable distribution of accidental errors to be expected from any number of similar random samples? In other words, how can we determine the unreliability of an average — its probable error — from a single sample, instead of by the use of the laborious method described, which is impossible in actual practice?

It can be demonstrated easily that *the less the variability of the individual values in a single representative sample, the less will be the variability in the means and other statistical measures obtained from many random samples.* This variability within the single sample is measured by the standard deviation of the distribution. Experiment shows that the more closely the size of the random sample approaches the total population, the closer the obtained average approaches to that of the entire population. *The variability of the mean and other statistical measures decreases as the size of the sample relative to the total population increases.* This decrease does not occur in direct proportion to the size of the sample, but *in proportion to the square root of the number of items in the sample.* This requires that in order to cut down the amount of variability by one half, not twice the number of items is required in the sample, but four times the number.

Therefore, *the error of the mean depends upon $\sqrt{N}$ and σ obtained from the single sample under investigation.* In actual practice in statistical work the sample is a given size governed by the judgment of the investigator and by his limitations as to time, money and available data. We now wish to know the unreliability of an average or a standard deviation computed from this sample which is assumed to represent the larger population.

FORMULÆ FOR COMPUTING THE UNRELIABILITY OF MEASURES OBTAINED FROM RANDOM SAMPLES

The formulæ in current use for measuring the variability of the mean and standard deviation are stated in terms of the number of items (N) and the measure of variability (σ) for the given sample under investigation. The derivation of these formulæ need not concern the student, since the terms used are familiar and the reason for their use has been explained.

The unreliability of a mean. Using the height data of the previous illustration, let us examine one random sample of 100 drawn from the larger population of 1000 heights. The mean of this single random sample is 67.53 inches, σ is 2.48 inches, and N is 100. It is the unreliability of 67.53 inches obtained from a single sample which is to be determined. The presumptive standard deviation (σ_M) of a large number of means obtained from random samples distributed about this most probable mean (67.53 inches) can be computed from the formula:

$$\sigma_M = \frac{\sigma_{\text{sample}}}{\sqrt{N}} = \frac{2.48 \text{ inches}}{\sqrt{100}} = \pm .25 \text{ inches}$$

The standard deviation of 100 actual sample means was ± .24 inches, almost the same value.

In the experiment with 100 samples of height it was shown that the mean of all the 100 means (67.50 inches) approximated very closely the mean of the entire population (67.57 inches). From that experiment it was also evident that a mean from any single sample chosen at random is more likely to fall at the mean of all the samples than at any other specified value. This is true because the means in that experiment were distributed about the mean of all the samples *as a mode*. Any random sample mean is more likely to fall at this modal value than at any other single value in the distribution.[1] Therefore, for purposes of describing the unreliability of the mean we regard the mean of the single sample (67.53 inches) as the central value in a distribution of other possible sample means.

Now σ_M (.25 inches) computed by the formula takes the place of the computation of many sample means and their variability, as was done in the actual height samples. It constitutes a measure of the *probable variability* of numerous sample means about the mean of the sample investigated (67.53 inches) without actually taking the samples and computing the separate measures. Having a measure of probable variability (.25 inches) it is easy to describe the expected distribution of numerous sample means about the value 67.53 inches as a mode, according to the probability curve.

[1] In describing the probable error of a mean or other statistical constant obtained from a single sample, the significant thing is the probable distribution of the means obtained from many similar samples. Our experiment with heights has shown the nature of this distribution about the mean of all the means as a mode. We were able to state the chances of any random sample mean falling within a given range plus and minus from the mean of all the samples. This divergence constitutes the error of the sample mean. Since the mean of the means is a mode about which other means are grouped in symmetrical form, the probable deviation of any random sample mean from this central value is less than its probable deviation from any other specified value. From our experiment we were able to state, in terms of the standard deviation of all the sample means, the *probable deviation* of any random sample mean from the mean of the means. This is the probable error of the mean of a single random sample.

Of other possible sample means 68.26 per cent may be expected to fall within the range of variation .25 inches plus and minus from 67.53 inches. It follows that the chances are 2 to 1 that any random sample mean will be included within this range of variation. The chances are 2 to 1, therefore, that a second sample mean taken at random will fall within the range 67.53 inches ± .25 inches, or between 67.28 and 67.78 inches. Likewise the chances are about 21 to 1 that a second sample mean will not differ from 67.53 inches by more than 2σ, or .50 inches plus or minus, a range extending from 67.03 to 68.03 inches. The chances of a deviation from 67.53 inches exceeding 3σ, .75 inches, are about 1 to 369. This amounts to *practical certainty* that any other sample mean will fall within .75 inches from 67.53 inches plus or minus, *so far as the accidental errors of sampling influence the results.* In this manner the unreliability of a mean is described and the limits of expected variation are defined and called the *standard error of the mean.*

The unreliability of a standard deviation. The measure of variability (σ) computed from a single random sample is itself a variable and subject to error. The probable variability of σ may be determined by the formula

$$\sigma_{\text{standard deviation}} = \frac{\sigma_{\text{sample}}}{\sqrt{2N}}$$

The absolute amount of variability or unreliability is naturally less for the standard deviation than for the mean. The $\sqrt{2N}$ replaces $\sqrt{N}$ in the formula used for the mean. For the sample of 100 heights

$$\sigma_{\text{standard deviation}} = \frac{2.48 \text{ inches}}{\sqrt{200}} = \pm .18 \text{ inches}$$

The explanation of the significance of $\sigma_{\text{standard deviation}}$ (± .18 inches) is similar to that already given for the variability of the mean. The standard deviation of the sample of heights should be expressed 2.48 ± .18 inches. This measures the *standard error* or *unreliability* of a standard deviation computed from a single sample.

The unreliability of a difference between two measures. Two averages may be computed each from a limited sample. It may be desirable to interpret the difference between these two averages M_1 and M_2. But each of them is a variable because computed from a random sample and, therefore, the difference is also a variable. The unreliability of a difference between the means M_1 and M_2 may be determined by the use of the formula

$$\sigma_{(M_1 - M_2)} = \sqrt{(\sigma_{M_1})^2 + (\sigma_{M_2})^2}$$

The unreliability of the difference between the mean M_1 and the mean M_2 is equal to the square root of the sum of the square of the unreliability of M_1 and the square of the unreliability of M_2. The result may be interpreted by the use of the probability curve in the same manner as the errors of the mean and the standard deviation.

The unreliability of a coefficient of correlation. A measure of the variability or unreliability of r, the Pearsonian coefficient of correlation, may be computed by the formula

$$\sigma_r = \frac{1 - r^2}{\sqrt{N}}$$

This measure will be treated in the following chapter.

Unreliability measured in terms of probable error. So far the variability or error of statistical measures describing their unreliability has been expressed as σ, the standard error. Another measure in current use is the probable error ($P.E.$) which was shown to be .6745 σ. This measure has a logical significance which has brought it into current use for expressing the unreliability of statistical measures. *It is the range of variation, measured plus and minus from the central value, within which 50 per cent of the values fall* (see Figure 27). It has the same meaning as the semi-interquartile range in a single distribution. The chances are equal that any value taken at random will fall within this range of variation above and below the mean or outside this range. Probable error should be used to describe the distribution of errors arising in the process of sampling and should not be applied to deviations occurring within a single frequency distribution.

The formulæ which have been given in terms of σ are now stated in terms of $P.E.$

$$P.E._M = .6745 \frac{\sigma_{\text{sample}}}{\sqrt{N}}$$

$$P.E._{\text{standard deviation}} = .6745 \frac{\sigma_{\text{sample}}}{\sqrt{2N}}$$

$$P.E._{(M_1 - M_2)} = .6745 \sqrt{(\sigma_{M_1})^2 + (\sigma_{M_2})^2}$$

$$P.E._r = .6745 \frac{1 - r^2}{\sqrt{N}}$$

The values of $P.E.$ are interpreted in the same manner as those of σ, by the use of the probability curve (Figure 27). The values of $P.E.$ in terms of σ are: 1 P.E. = .6745 σ; 2 $P.E.$ = 1.3490 σ; 3 $P.E.$ = 2.0235 σ; 4 $P.E.$ = 2.6980 σ. The table in Appendix D gives the percentages of the total area of the probability curve included within these distances in terms of σ from the maximum ordinate.

In the random sample of heights for which $\sigma_M = .25$ inches, the $P.E._M$ is .6745 times .25 inches, or .17 inches. The chances are even that a second sample mean will be located within a range of ± .17 inches from 67.53 inches, between 67.36 and 67.70 inches; the chances are $4\frac{1}{2}$ to 1 that another sample mean will fall within ± 2 *P.E.* (± .34 from 67.53 inches); the chances of falling within ± 3 *P.E.* (± .51 inches) are 22 to 1; and within ± 4 *P.E.* (± .68 inches) the chances are 142 to 1. Therefore, 67.53 ± .17 inches describes the unreliability of this mean obtained from a random sample, in terms of the probable error. *The probable error defines an interval, symmetrically including the computed mean, such that the chances are even that any other sample mean taken at random will fall within it.* The chances 142 to 1 indicate with practical certainty that another sample mean will not differ more than 4 *P.E.* from the computed value, 67.53 inches, so far as the divergence is due to the accidental conditions of sampling. We are practically certain that any other random sample mean will fall within the range 67.53 ± .68 inches, or 66.85 to 68.21 inches.

Degrees of probability. It may prove useful to the reader to summarize at this point the degrees of probability indicated by different amounts of σ and *P.E.* in defining the unreliability of statistical measures.

Amounts of σ	Chances of Occurrence	Amounts of *P.E.*	Chances of Occurrence	Degrees of Probability
$\pm .6745\,\sigma$	1 to 1	± 1 P.E.	1 to 1	Equal
$\pm 1\ \sigma$	2 to 1	± 2 P.E.	$4\frac{1}{2}$ to 1	Favorable
$\pm 2\ \sigma$	21 to 1	± 3 P.E.	22 to 1	High
$\pm 3\ \sigma$	369 to 1	± 4 P.E.	142 to 1	Practical Certainty

"Practical certainty" means that if values obtained from similar samples fall more than $\pm 3\,\sigma$ or ± 4 *P.E.* from the corresponding value computed from the first sample, the variations are almost certain to indicate *significant differences* in the phenomena under investigation and cannot be attributed to the errors of sampling.

GAINS IN WEIGHT — AN APPLICATION OF PROBABLE ERROR[1]

The object of this experiment was to determine the relative value of speltz and barley as a single grain ration for fattening sheep. Two lots of

[1] The facts are found in Bulletin 165, University of Illinois Agricultural Experiment Station (1913), pp. 478–79.

animals of 12 each were carefully selected and handled so as to render all conditions of the experiment uniform with the exception of the ration. Lot A was fed speltz and Lot B was fed barley for 105 days. The daily gain in weight was recorded for each animal. Uncontrolled conditions of the experiment, as feeding capacity, physiological peculiarities, activity, temperamental differences, caused variations in individual gains within each lot of 12 animals. The results were:

	Lot A	Lot B
Ration for 105 days	Speltz	Barley
Number of animals	12	12
Mean gain per animal	25.0 pounds	37.9 pounds
σ of gains	9.44 “	8.23 “
σ_M (unreliability)	2.73 “	2.38 “
$P.E._M$ (unreliability)	1.8 “	1.6 “
Restatement of mean gain	25.0 ± 1.8 “	37.9 ± 1.6 “

The expression for mean gain of Lot A, 25.0 ± 1.8 pounds, signifies that the chances are even that if the experiment were duplicated as closely as possible with 12 other animals, the mean gain would fall within the range 23.2 to 26.8 pounds inclusive. Furthermore, it is practically certain (odds 142 to 1) that a second experiment would show a mean gain within 25.0 ± (4 times 1.8) or between 17.8 and 32.2 pounds. In other words, 50 per cent of the mean gains obtained from a large number of similar experiments carefully controlled would not differ by more than 1.8 pounds from the mean (25.0 pounds) obtained from this first experiment, *so far as accidental, uncontrolled conditions influence the result.* This defines the unreliability of the mean gain obtained from the given experiment A.

Lot B showed a mean gain of 37.9 pounds, which is much greater than Lot A, but we have seen that a part of this difference may be due to accidental conditions and not to the barley ration, which is the only controlled condition causing a difference between the two lots of animals. What is the unreliability of the mean gain for Lot B? The chances are even that any second similar experiment with the barley ration will show a mean gain of 37.9 ± 1.6 pounds, or between 36.3 and 39.5 pounds. The odds are 142 to 1 that the mean gain of a second random experiment will fall between 37.9 ± (4 times 1.6 pounds), or 31.5 to 44.3 pounds. This defines the unreliability of the mean gain of Lot B.

We may feel sure that the *one controlled difference* in the treatment of Lots A and B, the difference in grain ration, does influence the gain in weight, barley tending to produce the better gain under the conditions of this experiment. But it is also clear that not all the difference between

25.0 and 37.9 pounds can be safely attributed to the different rations, because we have shown the *variability of each of these mean gains due to accidental, uncontrolled conditions.*

In Figure 29 the distribution of errors in the mean gains of Lots A and B according to the probability curve is shown in terms of both σ and probable error. The following tabular statement of the probable distribution of mean gains from other similar experiments will assist the student in understanding the diagram:

Lot A

Within a range of			Chances of Occurrence
±1 σ from 25.0	=25.0±2.73=22.27 to 27.73	pounds	2 to 1
±2 σ " "	=25.0±5.46=19.54 to 30.46	"	21 to 1
±3 σ " "	=25.0±8.19=16.81 to 33.19	"	369 to 1
and			
±1 P.E. " "	=25.0±1.8 =23.2 to 26.8	"	1 to 1
±2 P.E. " "	=25.0±3.6 =21.4 to 28.6	"	4½ to 1
±3 P.E. " "	=25.0±5.4 =19.6 to 30.4	"	22 to 1
±4 P.E. " "	=25.0±7.2 =17.8 to 32.2	"	142 to 1

Lot B

±1 σ from 37.9	=37.9±2.38=35.52 to 40.28	pounds	2 to 1
±2 σ " "	=37.9±4.76=33.14 to 42.66	"	21 to 1
±3 σ " "	=37.9±7.14=30.76 to 45.04	"	369 to 1
and			
±1 P.E. " "	=37.9±1.6 =36.3 to 39.5	"	1 to 1
±2 P.E. " "	=37.9±3.2 =34.7 to 41.1	"	4½ to 1
±3 P.E. " "	=37.9±4.8 =33.1 to 42.7	"	22 to 1
±4 P.E. " "	=37.9±6.4 =31.5 to 44.3	"	142 to 1

Figure 29 A and B shows the distributions of these accidental variations or errors in the two means on the hypothesis that many similar experiments with twelve animals each,[1] controlled in the same manner, have been made.

Finally, we can use our formula for the probable error of a difference to help us settle the question of the significance of the mean gains of Lots A and B. The computed difference is 37.9 pounds − 25.0 pounds = 12.9 pounds. The *probable error of this difference* is

$$\begin{aligned} P.E._{\text{difference}} &= .6745\sqrt{(2.73)^2 + (2.38)^2} \\ &= .6745\sqrt{13.1173} \\ &= 2.4 \text{ pounds} \end{aligned}$$

We may now state the unreliability of the difference as 12.9 ± 2.4 pounds. Even 4 *P.E.*, or 4 times 2.4 which equals 9.6 pounds, would not equal the observed difference, 12.9 pounds. Therefore, we are practically certain

[1] The number of animals is small and therefore the probable error is large. The difficulty of controlling the conditions is greatly increased with the size of the sample.

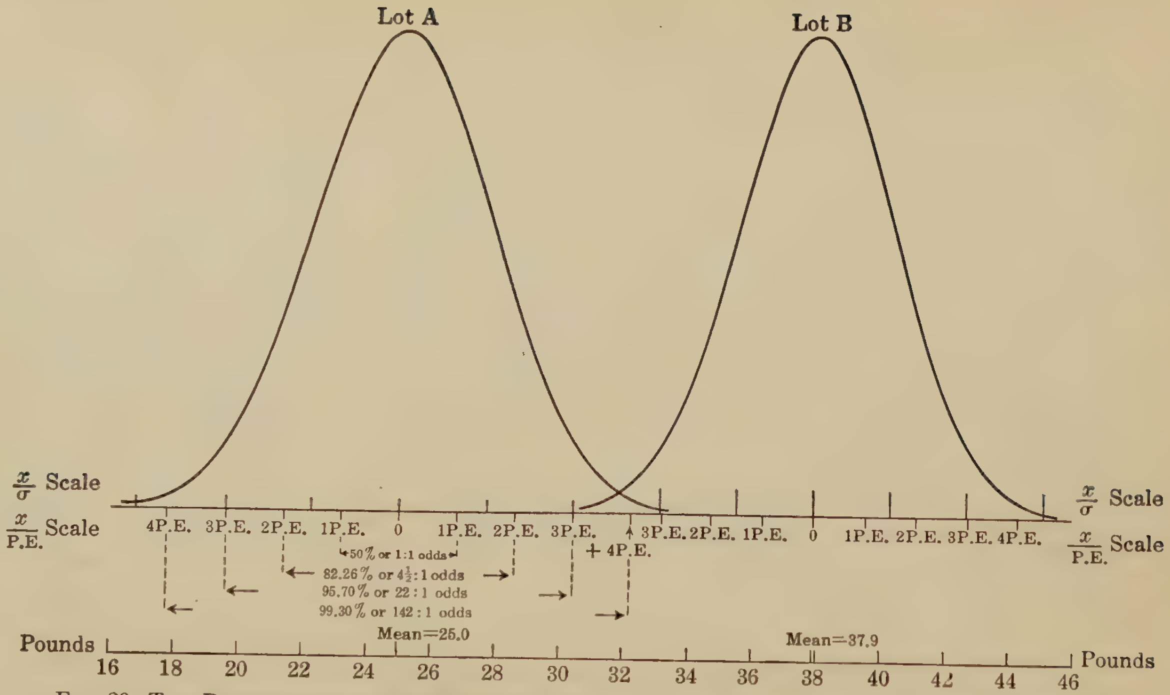


FIG. 29. THE DISTRIBUTIONS OF THE EXPECTED ERRORS OF THE MEAN GAINS IN WEIGHT IN FEEDING EXPERIMENTS

(The ration for Lot A was speltz and for Lot B was barley. Data from Bulletin 165, Agricultural Experiment Station, University of Illinois, July, 1913, pp. 478 and 479.)

that not all this difference could be due to the accidental variations in the two mean gains obtained from the two random experiments. *If we continued to experiment, this difference in favor of the barley ration would continue to appear as a true difference.*

PROBABLE ERROR AND REPRESENTATIVENESS OF A SAMPLE

When the probable error is small it does not prove that the sample is representative. Suppose that in an investigation of wages which is presumed to represent general wage conditions only union wage rates are collected by the sampling method. The number of cases may be large and the work on union wages may be carefully done. The probable error of the average computed by the formula $.6745 \frac{\sigma}{\sqrt{N}}$ may be small. Nevertheless, the sample does not represent general wage conditions. It represents only union wage conditions. There exists a *constant error* in making up the sample which should have been controlled by the investigator and which has nothing to do with the *size of the sample* and the accidental variations. The error will still persist regardless of the number of cases investigated. The sample may be *adequate* to take care of chance variations and to make the measures reliable in the sense discussed in this chapter and yet not be *representative.* Constant errors and bias of all kinds which affect the representativeness of a sample must be treated by different methods which are discussed in connection with the explanation of the procedure of sampling in Chapter XIV.

Probable error as a warning. If we take successive samples chosen in a similar manner from the same population, and we find repeatedly that the averages or other measures fall outside 4 *P.E.*, it is a *warning that other than accidental factors are causing variations.* The entire procedure of sampling should then be reviewed. So long as variations of that sort take place conclusions from the results are useless. *Conditions which should be controlled are probably being neglected, and these should be discovered.*

THE IMPORTANCE OF THE SIZE OF THE SAMPLE

Can we say that the sample must never have less than 100 cases, or that it will be adequate if it includes this number or more? The discussion of this chapter does not justify such an inference. The probable error of the mean depends upon both the standard deviation of the sample chosen and the number of cases included. If the data from which a small sample is taken at random are similar, in the formula $\frac{\sigma}{\sqrt{N}}$

the standard deviation will be small, but the number of cases is small also, and the resulting probable error will be relatively great as compared with that of a larger sample of the same data. If the data show wide variation the standard deviation for a few cases is likely to be very large due to the chances of selecting extreme variants, and since the number of cases is small the probable error will be very large. Our confidence in the mean or other statistical measures depends upon the size of the probable error relative to the size of the measure in question. The *minimum size of the sample* which should be chosen in a given situation *depends upon the homogeneity or similarity of the data from which it is selected and the standard of accuracy established for the problem under investigation.*

The probable error merely describes the amount of variation which may be expected in the statistical measures under the conditions of variation exhibited by the sample and with the number of cases investigated. Both these factors are given weight in the formulæ used. No general rule can be stated as to the minimum size of sample which will apply to all investigations. Each situation must be judged according to its own characteristics. *No summary statistical measure computed from a sample should be stated without defining, if possible, its probable variation due to the accidental conditions of the sampling procedure.*

READINGS

Rugg, H. O., *Statistical Methods Applied to Education*, chaps. 7 and 8.

Mills, F. C., *Statistical Methods as Applied to Economics and Business*, chap. 15 ("Elementary Probabilities and the Normal Curve of Error"), and chap. 16 ("Statistical Induction and the Problem of Sampling").

Pearl, Raymond, *Medical Biometry and Statistics*, chap. 10 ("Probable Error Concept." Good examples), and chap. 11 ("Elementary Theory of Probability").

Jerome, Harry, *Statistical Method*, chap. 10.

Thorndike, E. L., *An Introduction to the Theory of Mental and Social Measurements*, 2d ed., chap. 12 ("The Reliability of Measures").

King, W. I., *Elements of Statistical Method*, chap. 8 ("Approximation and Accuracy").

Weld, L. D., *Theory of Errors and Least Squares*. (Excellent introduction to measurement and the properties of errors, with examples.)

Whipple, G. C., *Vital Statistics*, 2d ed., chap. 13 ("Probability").

REFERENCES

Rietz, H. L., *Handbook of Mathematical Statistics*, chap. 5. (Pearson's test of goodness of fit, pp. 78–81. On page 77 is a convenient table of the Probable Errors of many statistical constants.)

Kelley, T. L., *Statistical Method*, chap. 5 ("Derivation of Equation of Normal Distribution").

Elderton, W. P., *Frequency Curves and Correlation*.

Yule, G. U., *An Introduction to the Theory of Statistics*, 6th ed., chap. 15 ("The Normal Curve"). (An excellent bibliography at the end of the chapter.)

Bowley, A. L., *Elements of Statistics*, 4th ed., Part II, chap. 2. (Part II is devoted to "Applications of Mathematics to Statistics.")

Jones, D. C., *A First Course in Statistics*, chaps. 12, 13 and 14 ("Probability and Sampling"), chap. 18 ("The Normal Curve of Error"). Intervening chapters devoted to the theory of curve fitting and applications.

Carver, H. C., *Handbook of Mathematical Statistics* (H. L. Rietz, Editor), chap. 7 ("Frequency Curves").

West, Carl J., *Introduction to Mathematical Statistics*, chap. 6 ("The Normal Probability Curve").

Whitaker, E. T., and Robinson, G., *The Calculus of Observations.*

Brunt, David, *The Combination of Observations.*

Griffin, F. L., *An Introduction to Mathematical Analysis.*

Keynes, J. M., *A Treatise on Probability.*

Coolidge, J. L., *An Introduction to Mathematical Probability*, Oxford University Press, London, 1925.

Pearson, Karl, *Tables for Statisticians and Biometricians* (Introduction).

[Details as to publisher and date of publication of the readings and references given at the close of each chapter are found in the alphabetical list in Appendix A.]

CHAPTER XII

DISCOVERY AND MEASUREMENT OF RELATIONSHIPS — CORRELATION

STATISTICAL analysis and interpretation usually require the investigation of associations between two or more series of facts. Previous chapters have explained and illustrated methods of classifying, summarizing and describing a single series of quantitative data by means of averages, measures of dispersion and graphic representation. The next problem in the scientific treatment of data is to discover significant relationships. As a rule phenomena are neither *absolutely independent* nor *absolutely dependent — they are associated in varying degree.* The problem is to determine in each case the degree of association, as this indicates the significance of the relationship. Is it possible to formulate a general statement or law which describes the relationship and which meets the test of new facts? How accurately will such a description of past experience enable us to predict future experience?

How are changes in money wages over a period of time related to changes in retail prices? The *real wages* of the worker depend upon this relationship. Are fluctuations in birth-rates and death-rates in the population associated with differences in economic status, and if so how closely? Does the infant death-rate fluctuate with the income of the family and with natural conditions of temperature and humidity? How do different series of quantitative data behave with respect to each other in the business field in times of prosperity and depression? Does a change in one series foreshadow variations in others, by regularly antedating them? Answers to these and scores of other questions require the investigation of relations between different groups of facts.

NATURE OF THE INQUIRY INTO CAUSE AND EFFECT RELATIONS[1]

It is one of the fundamental requirements of scientific method that the sequence of events be described and understood. Professor Pearson finds the essential idea of causation in a *routine of perceptions.* A certain

[1] Karl Pearson; *Grammar of Science*, chaps. IV and V. All students should read these chapters for the philosophy of contingency and correlation, and to Professor Pearson the author acknowledges his indebtedness. The reader should refer to Chapter III of the present volume for a review of the essentials of scientific method. It is not our purpose to trace historically the origin of this interpretation of causal relations.

association or sequence of phenomena has recurred again and again. We believe that the same association or sequence will recur in the future and we express this belief in our concept of *probability.* This belief that the future will resemble the past experience constitutes the basis for *prediction.* The faculty in man which enables him to place his sense impressions in some fairly constant order makes rational life possible.

For example, observation has shown that the amount of rainfall and the temperature at the critical stage of the growing crop influence the amount of the yield. Variations in these conditions have been associated in experience with fluctuations in the yield per acre. Knowing the antecedent conditions of weather enables the crop expert *to predict with more or less accuracy* the expected production. It must be emphasized, however, that there exists in this situation *no absolute and necessary inherent causal connection* between rainfall at the period of growth and the amount of the crop. Irrigation changes the conditions upon which the crop depends so as to render the two variables, rainfall and yield per acre, *largely independent of each other.* We can describe experience but we cannot arrive at an explanation which involves *necessity.* It is possible to describe it in terms of *probability.*

Furthermore, *routine of experience* does not usually connote *absolute sameness* but only a *degree of sameness.* It is a relative term because experience does not repeat itself with absolute exactness. For example, given weather conditions are associated with *about the same* yield per acre. In the preceding chapter it was noted that repeated measurements of the same object do not produce the same results. The conclusions of the scientist are based upon *average experiences,* eliminating the variations by the *concept of the most probable value.* Families of a given size and social status, having the same amount of income, do not spend exactly the same proportion of it for food, but *about the same.* We average the food expenditures of many families and state the result as a *general fact.* But *the average is an ideal concept derived from experience* which is significant only when considered along with the individual deviations from it. Likewise the *concept of causation* is drawn as an ideal from experience — *the reality is the routine of perceptions.*

In the sense just explained *anything is a cause which regularly antedates or accompanies a phenomenon.* If we vary one factor, to what extent is there an associated change in a second factor which is presumed to be related to the first? There may be no observed change, which indicates *independence.* If we should find that when boys of a certain age increase one inch in height they always increase five pounds in weight, we would say that there exists *absolute dependence* of weight and height. *Between* *these two limits of absolute independence and absolute dependence all degrees of association may happen.*

When one factor varies the other may fluctuate in sympathy, but not always to the same extent. For every value in the one series there usually results a *distribution of values* in the related series. For example, families having an income of $2000 spend varying proportions for food, ranging from 30 to 42 per cent. The less this variation in the related phenomenon, as food expenditures, the closer the *association or correlation.* If there were no variation in the proportion spent for food in a specific income group then food expenditure would be absolutely related to the size of income — *the correlation would be perfect.* If the variation in the related series, the proportions spent for food by various families having about the same income, is greater, the correlation is less significant. The degree of variation in the related series corresponding to a given value in the other series measures the closeness of the association. *Degrees of association between series characterize them as passing from independence to dependence or causal relationship.*

Therefore, we no longer search for cause and effect relations as fixed and unvarying laws. Association or correlation between occurrences tends to replace the older idea of causation in scientific investigation. We have seen that variation is a universal characteristic of phenomena. We can secure relative likeness in phenomena by a process of classification which places similar things together and disregards minor variations. The problem of science is to find out how the variation in one group of facts is associated with or contingent upon the variation in other groups, and to measure the degree of the association.

The aim is to find the series of facts which are most closely correlated in order to enable the investigator to predict future experience. *Causation becomes a descriptive concept reached by statistical processes applied to the facts of experience.* These statistical methods and processes constitute the subject-matter of the present chapter.

LIMITATIONS OF THE EXPERIMENTAL METHOD IN THE SOCIAL SCIENCES

The physical sciences have developed in the laboratory not only the means and methods of accurate observation and measurement, but also methods for the control of many of the variable factors. The investigator observes the result of allowing a particular factor to vary, while he keeps all other conditions of his experiment constant.

In the social sciences the investigator is dealing with human beings and their related activities. He cannot control the individual units as if they were chemicals in a test tube. His freedom in experimentation is greatly restricted. Human nature and social and business customs stand in the way. Furthermore, the factors which influence a specific situation or are associated with a particular occurrence are likely to be so numerous that it is impossible to keep all the chief variables constant except the specific factor under observation.

For example, the investigator may be concerned with the discovery of the relative importance of the various factors associated with a high death-rate among infants. It is hoped to decrease the abnormal mortality in certain areas or among certain groups by gradually changing specific conditions associated most closely with the high rates, but where should a beginning be mȧde? The hypothesis may be advanced that the large percentage of married women employed in industry is chiefly responsible in the community under observation. But in the same community is found also a condition of low wages, a large proportion of foreign-born population, a high degree of illiteracy, great crowding in living quarters, lax regulation of general sanitary conditions, and an inadequate protection of the milk supply. What factor is most important? Usually there is no means of changing a particular condition quickly, while other conditions remain constant, in order to observe the effects upon the health of infants. *Other things do not remain the same while a single factor changes.*

In the social sciences it is necessary to devise statistical methods for isolating certain factors while others are being observed, and for measuring the degrees of relationship existing between significant variables which can be objectively measured and classified but which cannot be controlled easily for purposes of experiment.

MEASURING DEGREES OF RELATIONSHIP

The reason for observing the fact and measuring the degree of association between series of data has been stated. The methods required differ according to the nature of the facts, the number of related factors and the type of association. Parts of the subject are beyond the scope of this text, for example, multiple and partial correlation and a comprehensive treatment of curve fitting. Therefore, we shall define briefly the limits of our present discussion.

As shown in Chapter IV, facts may be grouped in categories which are defined *qualitatively*,[1] or the original observations may be magnitudes

[1] For a discussion of Pearson's Coefficient of Mean Square Contingency, see H. O. Rugg: *Statistical Methods Applied to Education*, pp. 299–307; also G. U. Yule: *An Introduction to the Theory of Statistics* (6th ed., 1922), pp. 64–67.

varying in amount and forming *quantitative* series. *This discussion is concerned with the measurement of degrees of association between quantitative series.*

A series of quantitative data, as the general death-rates in different districts, will be found to be associated with more than one variable factor. The death-rate may be shown to vary in sympathy with the number of persons per room in a city population. But both crowded housing conditions and high death-rates may be closely associated with poverty. If housing conditions were improved without a change in the level of incomes, it might not prove very effective in lowering death-rates. Not only is each of these factors associated more or less closely with mortality changes, but they are not indèpendent of each other. While one factor varies the others do not usually remain the same. Some statistical device is required for keeping certain factors constant while the relationship between others is being measured. This problem is one of *multiple and partial correlation.* The present treatise deals only with the *correlation of two variables.*

Correlation involves measurement of multiple factors which are thought to influence a particular phenomenon or event. For example, in describing the physical condition of a school child we measure height, weight, and age. Having obtained measurements for many children we attempt to discover significant relationships between these factors. A family standard of living is described by data on the size of income, by the proportions spent for various purposes, and by the number of dependents. Having obtained these facts for many families, we study the relationships of the different factors. We describe business conditions by gathering and relating data on commodity prices, interest rates, volume of production and many other factors. *Our purpose in all these cases is to furnish a more complete explanation of a particular phenomenon by examining the relationships between the factors which are associated with it.*

The scientific investigator seeks an *hypothesis* which describes in the best possible manner the association between two series of data. In some cases the variations in one series corresponding with variations in another are best described in generalized form by a *straight line.* If this hypothesis proves to be the best description of the facts, we may call it a *law of association.* In other cases the correspondence is more completely represented by some other form of curve. *We shall begin with the linear or straight line type of association.*

Since special methods are required in the analysis of time series and in measuring the degree of association between the fluctuations of two such series, these methods will be explained and illustrated in Chapter XIII.

THE SCATTER DIAGRAM

A scatter diagram presents a graphic description of the quantitative relation between two series of facts. Suppose we choose at random one hundred freshmen from the population of one thousand already used for illustration. The purpose is to show how variations in the heights of these students are related to variations in their weights. Are lower heights associated with lower weights and greater heights with greater weights? If this relationship exists how rigid is the bond of association? When the height is one inch greater, is the weight of these individuals always more by a definite number of pounds? Or, does the weight sometimes remain about the same, regardless of a variation in height, and sometimes increase with increasing height? Is a tall person just as likely to be light as heavy? Figure 30 represents the facts for one hundred pairs of heights and weights.

In Figure 30 the horizontal scale represents height in inches and the

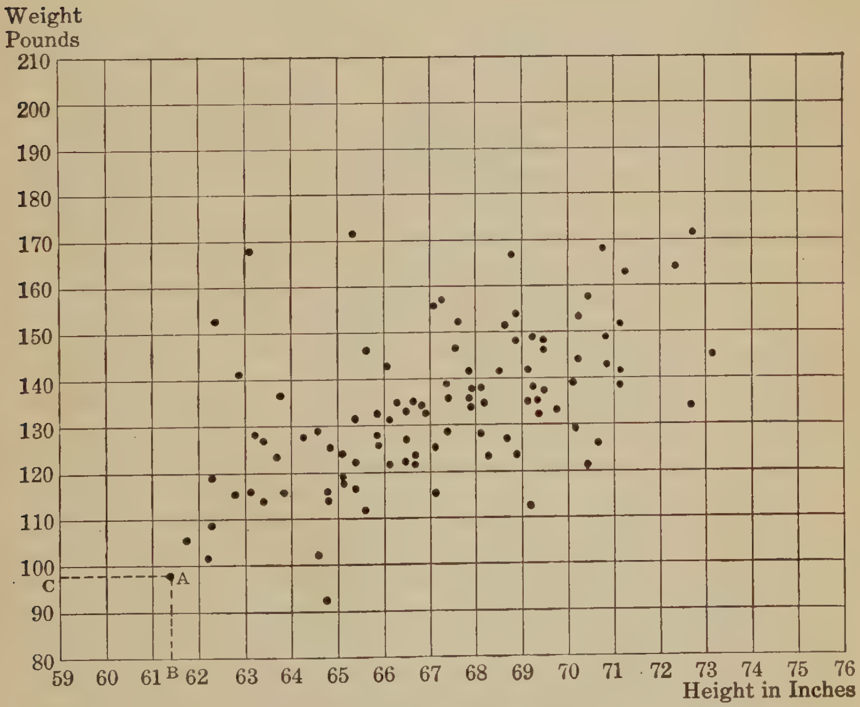


FIG. 30. A SCATTER DIAGRAM REPRESENTING THE QUANTITATIVE RELATION BETWEEN THE HEIGHTS AND WEIGHTS OF 100 INDIVIDUALS

(Data chosen at random from 1000 cases shown in Table 55, p. 295.)

vertical scale represents weight in pounds. Each pair of measurements, the height-weight for each individual, is shown by a dot so located on the diagram as to indicate the relation between the two facts for a given case. To locate any dot, A, the height of the individual is laid off on the horizontal scale and the weight on the vertical. Through these points, B and C on the respective scales, lines extend at right angles until they meet at A. The other dots are located in a similar manner. The number of dots can be easily increased to represent the entire one thousand cases.

Figure 30 indicates that, while there are a number of heavy individuals who are short in stature and others who are tall yet light in weight, *the tendency of the dots is to cluster along a diagonal trend* from the lower left hand corner toward the upper right hand corner of the diagram. In general this arrangement means that as height increases weight tends to increase also. If weight were independent of height, the dots would be found scattered over the page about the same average distance above the base line at each interval of height from left to right. If the association were absolutely rigid, then any given height would have only one value

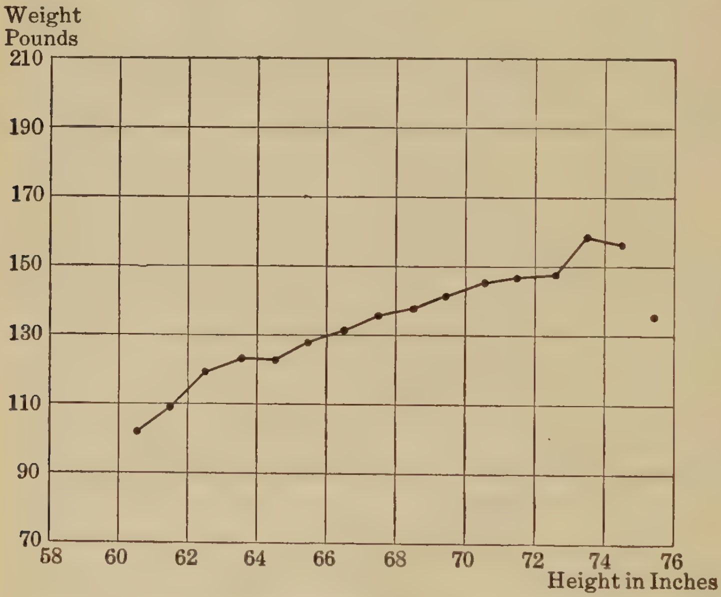


FIG. 31A. MEAN WEIGHTS RELATED TO GIVEN HEIGHTS — LINE OF MEAN WEIGHTS[1]

(Data from Table 55, p. 295, bottom row, means of columns.)

[1] The unconnected dot (.) in the diagram represents only a single case and is not significant.

for weight and the vertical scatter corresponding to any interval of height would be reduced to zero.

Actually, Figure 30 shows for the heights located between 62 and 63 inches *an array of corresponding weights,* and the dots are scattered vertically. This is true for successive intervals of height from left to right. *Therefore height and weight are neither absolutely independent nor absolutely dependent.* The degree of association ranges somewhere between these limits. How can the vertical and horizontal scatter, representing the actual measurements, be reduced to a form which will show the general trend more clearly?

Representation of vertical and horizontal distributions by averages. Individual dots on the scatter diagram representing the series of related measurements may mount into hundreds or thousands. It is too laborious to consider directly the more or less accidental position of each point. Yet an abbreviated description of these related values must not neglect the position of any dot.

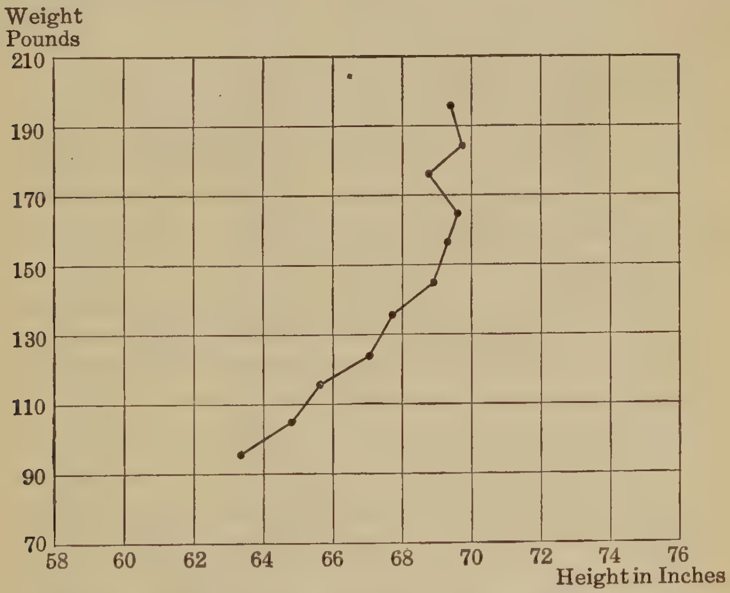


FIG. 31B. MEAN HEIGHTS RELATED TO GIVEN WEIGHTS — LINE OF MEAN HEIGHTS [1]

(Data from Table 55, p. 295, last column on right, means of rows.)

[1] The unconnected dot (.) in the diagram represents only a single case and is not significant.

We have already seen that an average is a value representing a group of detailed measurements. Therefore, the arithmetic averages of the several arrays of weights corresponding to given intervals of height *may be regarded as representing the respective arrays.* By substituting these *representative values* for the individual measurements of weight, the number of dots is greatly reduced. Actual heights are now related to the corresponding mean weights. The array opposite each interval of height is reduced to average weight, as shown in Figure 31A, page 254.

The direction of the change in weight associated with the change in height is clearly shown in Figure 31A. If height and weight were independent, all the dots would be arranged in random fashion around their average distance from the base line. The student, however, must not lose sight of the fact that *averages have been substituted for actual arrays of weight measurements,* because the scatter of the actual data from which the averages were computed has a significance to be explained later.

Likewise, any horizontal array of heights corresponding to a given interval of weight may be reduced to a mean height and may be plotted opposite the mid-value of the weight interval. Figure 31B, page 255, shows mean heights related to actual weights in this manner.

A DOUBLE-ENTRY TABLE OF RELATED VALUES — THE CORRELATION TABLE

We shall use for illustration the ages of husbands related to the ages of their wives. The association in this case is closer than between height and weight and it is clear that the relationship can be described in the best manner by the hypothesis of a straight line. Table 50 is a cross-classification in the form of frequency distributions in rows and columns of the ages of 5,317,520 pairs of husbands and wives who were residing together at the English census enumeration of 1901. The table is condensed by omitting 000's.[1] It shows how the ages of husbands are related to the ages of their wives, and *vice versa.*

The ages of husbands are grouped in five-year intervals in the horizontal direction and the ages of their wives in similar intervals vertically. For example, there are 4 husbands 15 and under 20 years of age, and the ages of their wives are distributed vertically in the third column from the left, 2 in the interval 15–20 and 2 in the interval 20–25 years. Likewise, there are 240 husbands 20 and less than 25 years of age, and the ages of their wives are distributed 1, 4, 46, 173, 16 in the next column to the right. Similarly each column of the table is a *frequency distribution* of wives' ages corresponding to the ages of husbands grouped in a five-year

[1] See Yule's *Introduction to the Theory of Statistics* (6th ed., 1922), p. 159.

TABLE 50. CORRELATION TABLE OR DOUBLE-ENTRY TABLE [a]

AGES OF HUSBANDS (X)

Ages of Wives (Y)	CLASS INTERVAL — 5 years ↓	f → ↓	15–	20–	25–	30–	35–	40–	45–	50–	55–	60–	65–	70–	75–	80–	85–	MEAN AGES OF HUSBANDS ↓
			4	240	688	817	793	700	595	483	369	277	175	104	50	18	4	
	85–	1														1		82.5
	80–	8											1	1	2	3	1	78.8
	75–	27										1	2	6	12	5	1	76.4
	70–	68									1	4	13	31	14	4	1	72.6
	65–	134							1	2	6	23	58	31	10	2	1	68.1
	60–	226						1	2	10	35	101	53	18	5	1		63.5
	55–	317					1	2	10	44	141	81	26	8	3	1		59.1
	50–	437				1	2	12	59	195	110	39	11	5	2	1		54.4
	45–	550			1	2	12	66	252	146	46	16	6	2	1			49.6
	40–	669			2	12	80	309	178	57	18	8	3	1	1			44.7
	35–	781		1	10	84	369	219	66	19	8	3	1	1				39.8
	30–	854		4	84	411	251	71	20	8	3	1	1					35.0
	25–	808		46	402	265	69	17	6	2	1							30.3
	20–	414	2	173	185	41	9	3	1									26.2
	15–	23	2	16	4	1												23.4
MEAN AGES OF WIVES →			20.0	23.4	26.9	31.1	35.6	40.2	44.9	49.5	54.0	58.4	62.6	66.4	69.3	73.3	75.0	

[a] The magnitude classes of the Y variable are arranged in order *from the bottom of the table toward the top*, which is contrary to the practice of some excellent authorities and to the prevailing practice in the case of a single frequency table. The author's chief reason for this arrangement is to have the table correspond to the graphic representation of correlation which is pictured in the straight line of generalized relationship — the so-called regression of Y on X (Figure 32). It seems important for mathematical as well as logical reasons that *positive correlation* should be indicated by a trend from the lower left-hand corner toward the upper right-hand corner of both the table and the diagram. In the first column on the left and in the first row at the top of the table only *the lower limits* of the class intervals are given.

class at the top of the respective columns. Or, there are 23 wives 15 and under 20 years of age. The ages of their husbands are distributed 2, 16, 4, 1 at successive intervals in the last row but one at the bottom of the table. Each row is a *frequency distribution* of husbands' ages corresponding to the ages of wives grouped in a five-year class at the left of the respective rows.

In tabulating the data for this table it is convenient to write the ages of both husband and wife on the same card in distinctive color of ink — a separate card for each pair. Then the cards may be sorted into five-year groups according to the ages of husbands. This sorting gives the *class-frequencies for all husbands to be entered at the top of the table in the second row.* Now the cards for each sub-group of husbands may be sorted again in five-year classes according to the ages of their wives. This sorting gives the frequencies to be entered in *each column of the table.* This procedure may be reversed and the first sorting may be made according to the ages of wives, which will give the *class-frequencies for all wives to be entered at each interval on the left of the table in the second column.* These sub-groups may be sorted again to secure the distributions of the ages of husbands to be entered in the *various rows of the table.*

It is clear that for a given age interval in the one series there is a distribution of ages in the other series. The columns distribute the ages of groups of wives and the rows distribute the ages of groups of husbands. Both rows and columns are sub-distributions of the total frequency. *Such a table sums up in its various compartments the items which would be represented by separate dots in a scatter diagram.* All cases in any compartment of the table are regarded as having the mid-value of the class-interval opposite which they are located, in contrast to the actual measurements represented on the scatter diagram. For any given age class of husbands, however, there still exists the scatter in the ages of their wives similar to the dots on the scatter diagram. The same is true for any given age class of wives.

Observation of the correlation table indicates a *positive association* between the ages of husbands and their wives. The association is designated *positive* or *direct* when variation in the one series is associated with variation in the same direction in other series. The correlation is called *negative* or *inverse* if the related variations take place in opposite directions. For example, an increase in the volume of business may be accompanied by a decrease in the amount of unemployment. The degree of association may be as high in negative as in positive association. As the age of husbands increases, there is a very constant tendency for the age of wives to increase also. The frequencies of the rows and the columns tend to mass in a diagonal upward direction across the page. The degree or closeness of the relationship has not been measured yet. Further summary and simplification is required.

Representation of the distributions in columns and rows by mean ages. In the bottom row of the table is found a series of *mean ages of wives* corresponding to actual ages of husbands as stated at the top of the table. Each column from left to right is treated separately as a frequency distribution of wives' ages. The five-year intervals on the left of the table are used for the computation of the means. Now the scatter in the columns has been summarized by *representative values* for the ages of wives related to the *actual ages of their husbands.* By comparing these two series of *actual* and *mean* values from left to right, the association between the ages of husbands and the ages of their wives may be more clearly shown. While the actual age of husbands varies from 15 to 90 years, the mean age of their wives varies from 20 to 75 years.

Or, the *actual ages of wives* on the left of the table may be related to the *mean ages of their husbands* stated in the extreme right hand column. These mean ages of husbands are computed in the same manner by regarding each row as a frequency distribution and using the five-year intervals at the top of the table for the computation. The scatter in the rows is now represented by the mean values for the ages of husbands. This enables us to state the range of variation in the *mean or representative values* of the one series associated with the range of variation in the *actual values* of the other series.

These means of columns and rows make more definite the observations made from the scatter diagram and from the correlation table, that direct association between the two variables is present. *The precise degree of association is still to be determined.* Some measure of degree is important in comparing one correlation with another, and to enable us to predict how much variation may be expected in a related series of values when we know the amount of variation in the given series and the closeness of the association between the two series.

THE GRAPHIC REPRESENTATION OF RELATIONSHIP

The dots extending diagonally upward across the page in Figure 32 describe the *association between the actual ages of husbands grouped in five-year intervals and the mean or representative ages of their wives.* Distances on the horizontal scale (the X axis) represent variations in the former, and on the vertical scale (the Y axis) represent variations in the latter.

The mean ages of wives, taken from Table 50, are plotted on the vertical scale above the corresponding mid-points of the five-year intervals representing the actual ages of husbands. The mean age of *all* husbands is 42.8 years and of *all* wives is 40.6 years. The line MO_1 cuts the X axis at the mean age of husbands and is drawn parallel to the Y axis. The line M_1O_1 cuts the Y axis at the mean age of wives and is parallel to the X axis. These lines cross at O_1, *the center of the system of related values.* From this center, regarded as zero, deviations may be measured in either the vertical or the horizontal direction. A zero point established in this manner is convenient, because frequently *the scales of the diagram do not start at zero.* In fact, there are no values below 15 years in the data presented, and in the height-weight problem no measurements were found less than 60 inches, and less than 90 pounds.

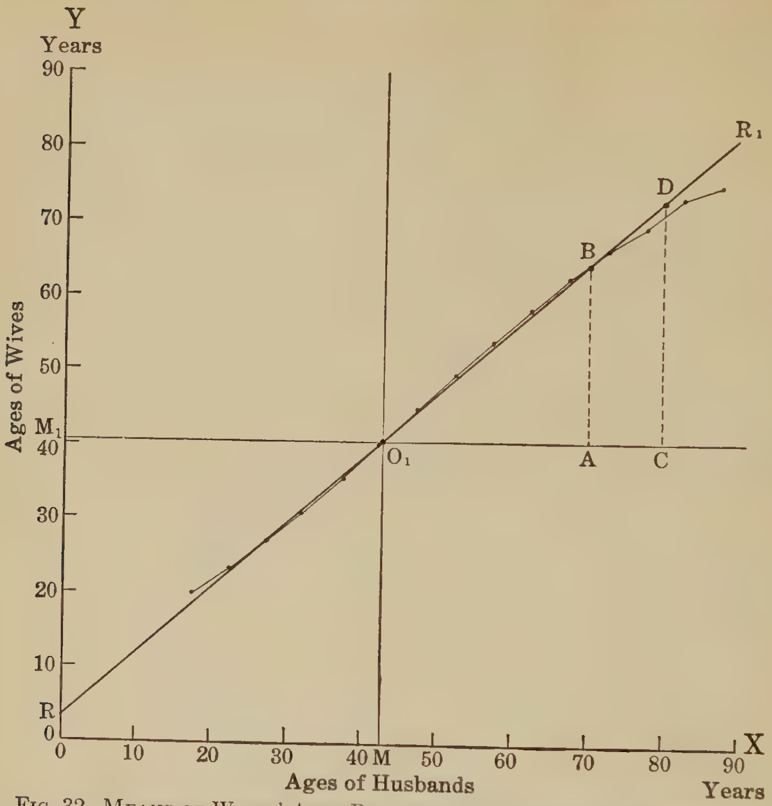


FIG. 32. MEANS OF WIVES' AGES RELATED TO GIVEN AGES OF HUSBANDS

The line of average relationship (regression of Y on X), RR_1, fitted by inspection to the means, on the linear hypothesis. (Data from Table 50. Equation to line RR_1 is $y = + .87\,x$; and the slope of RR_1 is $\frac{y}{x} = + .87$, with origin at 0_1.)

Now that the *representative values* are substituted for the *individual values*, what is the law of association of the representative points? It is clear from the positions of the dots that as the ages of husbands increase the mean ages of their wives also increase in *close positive association*, but how close is the association? We wish to obtain a quantitative index of the degree. The trend of the dots approximates closely a *straight line*, RR_1. If we adopt the *linear hypothesis* to describe the trend of the dots — the relationship — how may such a line be located and what does it mean?

The equation of a straight line passing through zero origin. The reader should refer to the description of a straight line and its graphic representation in Chapter XI (page 218). The general equation was stated $Y = mX + b$, in which b is a constant for the particular straight line, the distance from zero to the point where the line cuts the vertical axis; X and Y are values on the horizontal and vertical scales respectively, located with reference to the zero origin; and m describes the slope of the particular line under consideration.

If the straight line passes through the zero origin in Figure 25, b equals 0, and the equation becomes $Y = mX$. Now, any value (X) on the horizontal scale may be regarded as a deviation from zero and may be designated x; likewise, any value (Y) on the vertical scale may be regarded as a deviation from zero and may be designated y. Then, the equation of a straight line passing through zero origin may be stated $y = mx$. The slope of such a line is determined by the equation $m = \frac{y}{x}$, in which y is any vertical distance from zero, and x is any horizontal distance from zero. *It is characteristic of a straight line that for every point located on it the ratio* $\frac{y}{x}$ *has the same value.* This ratio is the inclination of the line — *its slope* — and is the tangent of the angle which the line forms with the horizontal axis.

In the correlation diagram (Figure 32) we have established a zero origin at the point where the lines of the means cross, *the center of the system of related values.* All horizontal and vertical deviations are measured from this point and are designated x and y respectively. Our purpose is to fit a straight line passing through this point as closely as possible to the means of all the columns of the correlation table.

Fitting the straight line in the correlation diagram. In Figure 32 the line RR_1 is fitted to the means of the columns of the correlation table *by inspection.* A transparent ruler or a thread and thumb tacks may be used for this purpose. This line can be fitted by more exact methods to be described later. It serves the purpose of a *generalized description of the trend of the points.* These points representing means are determined from the sub-distributions of ages by giving due influence to the frequencies in each compartment of the table. The slope of the line is described by the equation

$$m_1 = \frac{y}{x} = \frac{AB}{O_1A} = \frac{CD}{O_1C} = \frac{23.7 \text{ years}}{27.2 \text{ years}} = \frac{32.4 \text{ years}}{37.2 \text{ years}} = .87$$

Whether the distances AB and O_1A be measured with a ruler in inches or on the respective scales in years, the equation $\frac{AB}{O_1A} = .87$ is true. The value of m_1 may be substituted in the equation of the straight line $y = mx$, and we have $y = .87\,x$. This equation, $y = .87\,x$, enables us, if we know the value (x), of any horizontal deviation from O_1, to substitute it in the equation and to obtain a corresponding value for y. For example, in the diagram, O_1A = 27.2 years = x. Therefore, y = .87 times 27.2 years = 23.7 years = AB. The .87 in the equation is the ratio of any value of y to the corresponding value of x, always measuring these values from the center of the system of related values (O_1).

It must be remembered that x and y are deviations from the respective means of the related series regarded as origins; while X and Y are values at specific points on the respective scales laid off from zero origin. For example, O_1A = 27.2 years, measured on the horizontal scale from the mean, 42.8 years, to 70 years (70 − 42.8 = 27.2 years). Likewise, AB = 23.7 years, measured vertically from the mean, 40.6 years, to 64.3 years (64.3 − 40.6 = 23.7 years). In each case it is *variation from the mean which is measured and designated x and y*. The point A, however, is located at 70 years on the horizontal scale, and the point B at 64.3 years on the vertical scale. These values are designated X and Y.

Measurement of the degree of association. It is clear from the inclination of the line RR_1 that as the age of husbands increases the mean age of their wives also increases, but when we attempt to compare the amount of the change in one series with the amount of change in the other a difficulty appears. How much of a change in the age of wives is to be compared with a given variation in the age of husbands? *As the measurements are represented on the diagram they are not in comparable form.* Each series deviates from its own mean in a characteristic manner and the variation is measured by the *standard deviation* expressed in the unit of actual measurement (years). In this particular problem the units of both series are years, and the standard deviation of the one differs but little from the other in amount. Therefore, the variations in the two related series are *almost* comparable directly as represented in the diagram.

In other correlations the situation may be very different. For example, in the height-weight diagram (Figure 31A), the actual heights of one thousand individuals are associated with the mean weights of the same persons. How much variation in weight is to be compared with a given change in height? *Both the units of the two series and their standard deviations are widely different.* Height is expressed in *inches* and weight in *pounds*. The σ of one thousand heights is 2.6 inches and the σ of the weights is 17.0 pounds. How can we compare variations expressed in such different units and on such widely different scales? The same difficulty arises in comparing variations in the size of family incomes and the percentage of total income spent for food. Income is expressed in *dollars* and the related series in *per cents.*

Since it is associated variations which are being compared, it is clear from the examples cited that a given distance on the horizontal scale representing a variation measured in units of the one series may have a very different significance from the same distance on the vertical scale expressed in units of the other series. In other words, the ratio $\frac{y}{x}$, which determines the slope of the fitted line RR_1 in the diagram, does not measure the degree of association until the y and x are treated so as to make them comparable. *The problem is to express the actual variations in such series in terms of some measure common to both.*

The standard deviation — a common denominator. The tendency or capacity of each series to vary about the mean is measured by its standard deviation. *The actual amount of variation in any series is significant for comparison with another series only when expressed in terms of its own capacity to vary, its standard deviation.*

Let us apply this principle to the variations in the ages of husbands and their wives. The standard deviation of the entire distribution of the ages of wives, computed from the correlation table, is 12.7 years, and for the ages of husbands 13.1 years. The deviations shown in the correlation diagram may be expressed in terms of the respective standard deviations of each series.

$$\text{Now } \frac{y}{x} \text{ or } \frac{AB}{O_1A} \text{ becomes } \frac{AB \div \sigma \text{ Wives}}{O_1A \div \sigma \text{ Husbands}} = \frac{23.7 \text{ years} \div 12.7}{27.2 \text{ years} \div 13.1} = \frac{1.87\sigma^*}{2.08\sigma} = .9$$

$$\text{Likewise, } \frac{CD}{O_1C} \text{ becomes } \frac{CD \div \sigma \text{ Wives}}{O_1C \div \sigma \text{ Husbands}} = \frac{32.4 \text{ years} \div 12.7}{37.2 \text{ years} \div 13.1} = \frac{2.55\sigma}{2.84\sigma} = .9$$

The same procedure may be followed for the coördinates y and x of any point on the line RR_1. By the use of the respective standard deviations as units of variation, any change in the age of husbands on the horizontal scale is rendered comparable with the corresponding variation in the ages of wives on the vertical scale because both scales for measuring variations are transformed into *units of standard deviation.* It is really units of standard deviation which are related in the variations of the two series of values. Now the ratio $\frac{y \div \sigma_Y}{x \div \sigma_X}$ becomes *a measure of the degree of relationship and may be called the coefficient of correlation,* (r). The

[*] Both numerator and denominator are now expressed in units of standard deviation and are rendered comparable.

characteristics of the straight line are preserved since the same ratio (.9) is obtained for any point on RR_1, as illustrated above. When the numerator and denominator of the fraction from which the coefficient is obtained are equal, the association is complete and the *coefficient of correlation is unity.* This means that for a given variation in the one series there is *always* a proportional change in the corresponding value of the other series. The degree of association is between zero and unity and the coefficient is an index of this degree.

A more complete description of the line of relationship (RR₁). How may the line RR_1 in the diagram be described? Its simple equation is (1) $y = m_1x$, or $y = .87\,x$. In terms of the explanation given this equation becomes,

$$\frac{y}{\sigma_Y} = r\frac{x}{\sigma_X} = .9\frac{x}{\sigma_X}.$$

If $\frac{y}{\sigma_Y} = r\frac{x}{\sigma_X}$, multiplying both sides of the equation by σ_Y, then,

(2) $y = r\frac{\sigma_Y}{\sigma_X}x$, *which describes the line* RR_1. That (2) is equivalent to (1) is shown by substituting values for r, σ_Y and σ_X. We have, $y = .9\left(\frac{12.7 \text{ years}}{13.1 \text{ years}}\right)x$, or $y = .87\,x$. It will be seen that m_1 of equation (1) is equal to $r\frac{\sigma_Y}{\sigma_X}$ in equation (2). Both expressions describe the slope of the line RR_1.

The equation means that for every unit variation in x (age of husbands), we find .87 as much variation, *on the average*, in y (age of wives). It must not be forgotten that this relation is based upon *generalized experience* as represented in the straight line fitted to the means, which in turn summarize the scatter in the columns of the correlation table. Any such estimate of values for y from known values for x by the use of these equations *can be accurate only within certain limits determined by the scatter about the line of generalized relationship* RR_1. This point will be considered later in this chapter.

The line RR_1 is commonly called the *line of regression* of Y on X, a term applied by Galton in relating the stature of children to that of their parents. He spoke of hereditary characters as tending to "regress back toward the mean of the race." In economics and social science there is no reason for speaking of this line of the means of the correlation table as a regression line. It seems more appropriate to call it the *line of generalized or average relationship.*

Another form of this equation describing the line RR_1 will be readily understood from Figure 32 and its explanation. Small x and y in the preceding discussion have always referred to *deviations* measured from the means, distances rather than points on the scale. Let large X and Y designate any specific values at *definite points* on the scale. Then x becomes $(X - \overline{X})$ and y becomes $(Y - \overline{Y})$, the bars above the letters denoting the means of the series.

Now equation (2) becomes,

$$Y - \overline{Y} = r\frac{\sigma_Y}{\sigma_X}(X - \overline{X}) \qquad {}^*(3)$$

Both equations mean the same thing, (3) using the absolute values on the scale, and (2) the deviations from the means.

A second line of the means of the correlation table. In Figure 33 the *means of the rows* in the correlation table, the ages of husbands, are related to the *actual ages of wives*. The mean ages of husbands are plotted opposite the mid-values of the five-year intervals on the Y axis. The trend of the resulting dots represents the association *in generalized form* between the actual ages of wives and the mean ages of their husbands. The straight line CC_1 is fitted *by inspection.*

This line CC_1 is commonly designated the regression line of X on Y. It is drawn through the center at the crossing of the means of the system. Its slope is determined by $\frac{x}{y}$, the tangent to the angle EO_1C_1 which the line makes with the Y axis.

We are now relating variations (x) in the mean ages of husbands to variations (y) in the actual ages of wives, always measuring them from the center O_1. As before, the variations in each series are set forth in *units of their respective standard deviations.* Then,

$$\frac{x}{y} \text{ or } \frac{AB}{O_1A} \text{ becomes } \frac{AB \div \sigma \text{ Husbands}}{O_1A \div \sigma \text{ Wives}} = \frac{27.3 \text{ years} \div 13.1}{29.4 \text{ years} \div 12.7} = \frac{2.08\sigma}{2.31\sigma} = .9$$

$$\text{and } \frac{ED}{O_1E} \text{ becomes } \frac{ED \div \sigma \text{ Husbands}}{O_1E \div \sigma \text{ Wives}} = \frac{36.6 \text{ years} \div 13.1}{39.4 \text{ years} \div 12.7} = \frac{2.79\sigma}{3.10\sigma} = .9$$

The coefficient, r, equals + .9. The ratio $\frac{x}{y}$ for all points on the line CC_1, whose coördinates x and y are expressed as deviations in units of their respective standard deviations, produces the same coefficient.

* This equation of the straight line RR_1 may be stated in the form presented in Chapter XI, p. 218, $Y = m_1X + b_1$. Transposing $\overline{Y}$ in (3) we have,

$$Y = r\frac{\sigma_Y}{\sigma_X}X + \left(\overline{Y} - r\frac{\sigma_Y}{\sigma_X}\overline{X}\right), \text{ in which } r\frac{\sigma_Y}{\sigma_X} = m_1 \text{ and } \left(\overline{Y} - r\frac{\sigma_Y}{\sigma_X}\overline{X}\right) = b_1.$$

The *degree of association* has been measured by relating either series to the mean values of the other. The coefficient r is the same in both cases. *There are two regression lines or lines of generalized relationship* for each correlation table, with their appropriate equations. These lines RR_1 and CC_1 may be drawn upon the same diagram, and it is suggested that the student do so.

The line CC_1, Figure 33, may be described by either of two equations, (2) or (3) below, involving the measure of association (r) and the related standard deviations, similar to those describing RR_1 in Figure 32.

(1) $x = m_2 y$; or setting forth x and y in terms of their respective standard deviations,

$$\frac{x}{\sigma_X} = r\frac{y}{\sigma_Y}$$

Multiplying both sides of the equation by σ_X,

*(2) $$x = r\frac{\sigma_X}{\sigma_Y}y.$$ Substituting large X and Y,

(3) $$X - \overline{X} = r\frac{\sigma_X}{\sigma_Y}(Y - \overline{Y})$$

These equations, (2) or (3), mean that for a given variation in y (age of wives) the average variation in x (ages of husbands) will be $r\frac{\sigma_X}{\sigma_Y}$ as much. Substituting values in (2), $.9\left(\frac{13.1}{12.7}\right)$ for $r\frac{\sigma_X}{\sigma_Y}$ we have .93. Therefore, $x = .93\,y$. This ratio of x to y determines the slope of the line CC_1, which may be tested by dividing any x distance on the diagram, as AB, by the corresponding y distance O_1A, that is $\frac{27.3 \text{ years}}{29.4 \text{ years}} = .93$.

COMPUTATION OF r FROM THE CORRELATION TABLE — PEARSON'S FORMULA

We have seen that the means of the columns and of the rows of a correlation table summarize the facts by representative values in a form convenient for fitting the straight lines which describe the generalized

* It is seen that m_2 of equation (1) is equal to $r\frac{\sigma_X}{\sigma_Y}$ of equation (2). Both expressions describe the slope of the line CC_1. Equation (3) may be expressed in the form $X = m_2 Y + b_2$, presented in Chapter XI. Transposing $\overline{X}$ of (3) we have,

$$X = r\frac{\sigma_X}{\sigma_Y}Y + \left(\overline{X} - r\frac{\sigma_X}{\sigma_Y}\overline{Y}\right), \text{ in which } r\frac{\sigma_X}{\sigma_Y} = m_2, \text{ and } \left(\overline{X} - r\frac{\sigma_X}{\sigma_Y}\overline{Y}\right) = b_2.$$

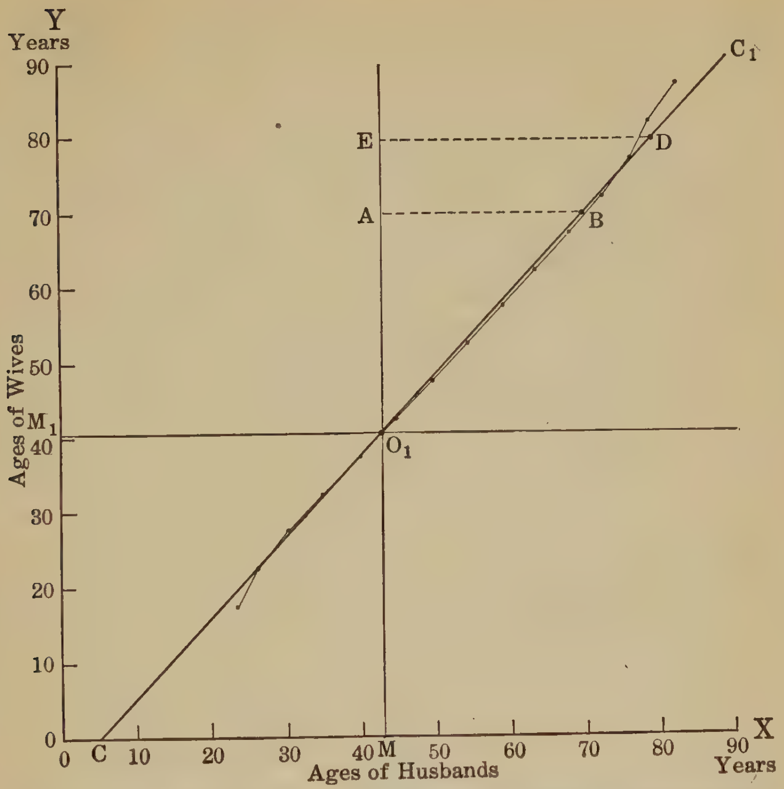


FIG. 33. MEANS OF HUSBANDS' AGES RELATED TO GIVEN AGES OF WIVES

The line of average relationship (regression of X on Y), CC_1, fitted by inspection to the means, on the linear hypothesis. (Data from Table 50. Equation to line CC_1 is $x = + .93\, y$; and the slope of CC_1 is $\frac{x}{y} = + .93$, with origin at 0_1.)

relationship. In the preceding explanation r was obtained by an approximate method and the straight lines were fitted by inspection.

In the illustration employed the means fall close to a straight line, the movements of the two variables are closely associated, the correlation is high, and the method of fitting by inspection proves very accurate, as will appear when tested by a more exact method. The inspection method becomes less accurate as the means fall less regularly along a definite trend. But we have seen the nature of the problem and we know what measures are required in order to state the equations which describe the line of generalized relationship fitted to the means of the rows or columns.

It has been emphasized in the preceding discussion that variations in two series cannot be related except in terms of units of their respective standard deviations. Hence, any formula measuring the degree of association must state the variations in units of standard deviation. Furthermore, it has been made clear that the means of the rows and of the columns summarize the facts of the sub-distributions in the correlation table, giving proper weight to the frequencies in each compartment or area. In 1896 Pearson published his product-moment or product-deviation method of measuring the degree of association, giving us the formula for computing the coefficient of correlation (r) and the equations of the lines of regression illustrated in the preceding pages.

Pearson's product-deviation method takes account of the corresponding x and y deviations of each compartment of the correlation table measured from the mean of the X series and the mean of the Y series, preserving the proper signs of the x's and y's, and using as weights the frequencies in each compartment of the table. The formula, as commonly stated, is

$$r = \frac{\Sigma xy}{N\sigma_X\sigma_Y}$$

in which the x's and y's are the deviations from the respective means of the series for each compartment of the correlation table, σ_X and σ_Y are the standard deviations of the entire X and Y distributions respectively, and N is the total number of related pairs of items.

The student is already familiar with the *short method* of computing the mean and standard deviation in a single distribution. We shall use this method also in computing r. The x and y deviations are expressed in *intervals or steps* from *guessed averages*, taken at the mid-value of a specific interval for the X and for the Y series (Table 51). Since r in the Pearson formula is a *ratio*, both the product-deviations of the numerator and the σ's of the denominator may be expressed in intervals or steps *throughout the computation.* The formula $r = \frac{\Sigma xy}{N\sigma_X\sigma_Y}$ may be restated,

$$*r = \frac{\frac{\Sigma d_X d_Y}{N} - c_X c_Y}{\sigma_X \sigma_Y}$$

* To show that the formulæ are equivalent, the following proof is presented:

Let d_{X_1}, d_{X_2}, etc. = the deviations from $G.A._X$

d_{Y_1}, d_{Y_2}. etc. = the deviations from $G.A._Y$

Then $d_{X_1} = x_1 + c_X$ and $d_{Y_1} = y_1 + c_Y$

$d_{X_2} = x_2 + c_X$ $d_{Y_2} = y_2 + c_Y$

etc. etc.

(c_X and c_Y being constants for the X and Y series.)

and $d_{X_1}d_{Y_1} = (x_1 + c_X)(y_1 + c_Y) = x_1y_1 + x_1c_Y + y_1c_X + c_Xc_Y$,

and $d_{X_2}d_{Y_2} = x_2y_2 + x_2c_Y + y_2c_X + c_Xc_Y$

etc.

in which the symbols d_X and d_Y, and c_X and c_Y have the same meaning as in Chapters VI and IX, but are applied to two series instead of a single series. They represent the deviations in intervals from the guessed averages of the X and the Y series respectively, and the correction factors in intervals for the guessed averages of these series. We have a d_X and a d_Y for each compartment of the correlation table. Every deviation from $G.A._X$ is in error by the amount of c_X, and every deviation from $G.A._Y$ is in error by the amount of c_Y, expressed in intervals. Therefore, c_X and c_Y are *constants* for their respective distributions.

Table 51 illustrates the procedure for correlating the ages of husbands and wives. Row (1) and column (A) give the mid-values of the five-year intervals of age. The guessed average for each series is chosen at 42.5 years. Any other mid-value may be selected for checking the computations, and it need not be the same for both series. Row (2) and column (B) state the respective frequencies for the entire X and Y distributions. Row (3) and column (C) give the deviations in intervals of each compartment of the table from the guessed average of the X series and from the guessed average of the Y series respectively, with the appropriate signs. All minus deviations are designated (−), all others are positive. The column and the row in which the guessed averages fall are enclosed by double rulings, and for these compartments of the table the deviations are zero. The *fd* products of row (4) are obtained by multiplying the corresponding items of rows (2) and (3) for each class, preserving the signs; likewise, the *fd* products of column (D) are obtained by multiplying the items of columns (B) and (C). These *fd* products are necessary for computing c_X and c_Y, the respective correction factors. Row (5) gives the fd^2 products obtained by multiplying the corresponding items of rows (3) and (4) for each class, all signs becoming positive; likewise, column (E) gives the fd^2 products obtained by multiplying

Summing, we have $\Sigma d_X d_Y = \Sigma xy + c_Y \Sigma x + c_X \Sigma y + \Sigma c_X c_Y$

But $\Sigma x = 0$, and $\Sigma y = 0$, and $\Sigma c_X c_Y = N c_X c_Y$

Substituting, the two middle terms disappear.

Therefore, $\Sigma d_X d_Y = \Sigma xy + N c_X c_Y$

Transposing, $\Sigma xy = \Sigma d_X d_Y - N c_X c_Y$

Dividing both sides by $N\sigma_X\sigma_Y$, we have

$$\frac{\Sigma xy}{N\sigma_X\sigma_Y} = \frac{\Sigma d_X d_Y - N c_X c_Y}{N\sigma_X\sigma_Y} = \frac{\dfrac{\Sigma d_X d_Y}{N} - \dfrac{N c_X c_Y}{N}}{\sigma_X\sigma_Y}$$

$$= \frac{\dfrac{\Sigma d_X d_Y}{N} - c_X c_Y}{\sigma_X\sigma_Y} = \text{Short Method Formula}$$

Table 51. Correlation between Age of Husband and Age of Wife — Short Method

	(A)	(B)	(C)	(D)	(E)	Ages of Husbands (*X*)															
(1)	*m*					17.5	22.5	27.5	32.5	37.5	42.5	47.5	52.5	57.5	62.5	67.5	72.5	77.5	82.5	87.5	
(2)		*f*				4	240	688	817	793	700	595	483	369	277	175	104	50	18	4	**5317**
(3)			*d* → *X*			−5	−4	−3	−2	−1	0	1	2	3	4	5	6	7	8	9	
(4)			↓ *Y*	*fd*		−20	−960	−2064	−1634	−793	**−5471**	595	966	1107	1108	875	624	350	144	36	**+5805**
(5)					*fd*²	100	3840	6192	3268	793		595	1932	3321	4432	4375	3744	2450	1152	324	**36518**
Ages of Wives (*Y*)	87.5	1	9	9	81														1 **72** *72*		
	82.5	8	8	64	512											1 **40** *40*	1 **48** *48*	2 **56** *112*	3 **64** *192*	1 **72** *72*	
	77.5	27	7	189	1323										1 **28** *28*	2 **35** *70*	6 **42** *252*	12 **49** *588*	5 **56** *280*	1 **63** *63*	
	72.5	68	6	408	2448									1 **18** *18*	4 **24** *96*	13 **30** *390*	31 **36** *1116*	14 **42** *588*	4 **48** *192*	1 **54** *54*	
	67.5	134	5	670	3350							1 **5** *5*	2 **10** *20*	6 **15** *90*	23 **20** *460*	58 **25** *1450*	31 **30** *930*	10 **35** *350*	2 **40** *80*	1 **45** *45*	
	62.5	226	4	904	3616						1 0	2 **4** *8*	10 **8** *80*	35 **12** *420*	101 **16** *1616*	53 **20** *1060*	18 **24** *432*	5 **28** *140*	1 **32** *32*		

Ages of Wives (Y)

57.5	317	3	951	2853					1 −3 *−3*	2 0	10 **3** *30*	44 **6** *264*	141 **9** *1269*	81 **12** *972*	26 **15** *390*	8 **18** *144*	3 **21** *63*	1 **24** *24*		
52.5	437	2	874	1748				1 **−4** *−4*	2 **−2** *−4*	12 0	59 **2** *118*	195 **4** *780*	110 **6** *660*	39 **8** *312*	11 **10** *110*	5 **12** *60*	2 **14** *28*	1 **16** *16*		
47.5	550	1	550	550			1 **−3** *−3*	2 **−2** *−4*	12 **−1** *−12*	66 0	252 **1** *252*	146 **2** *292*	46 **3** *138*	16 **4** *64*	6 **5** *30*	2 **6** *12*	1 **7** *7*			
42.5	669	0	**+4619**				2 0	12 0	80 0	309 0	178 0	57 0	18 0	8 0	3 0	1 0	1 0			
37.5	781	−1	−781	781		1 **4** *4*	10 **3** *30*	84 **2** *168*	369 **1** *369*	219 0	66 **−1** *−66*	19 **−2** *−38*	8 **−3** *−24*	3 **−4** *−12*	1 **−5** *−5*	1 **−6** *−6*				
32.5	854	−2	−1708	3416		4 **8** *32*	84 **6** *504*	411 **4** *1644*	251 **2** *502*	71 0	20 **−2** *−40*	8 **−4** *−32*	3 **−6** *−18*	1 **−8** *−8*	1 **−10** *−10*					
27.5	808	−3	−2424	7272		46 **12** *552*	402 **9** *3618*	265 **6** *1590*	69 **3** *207*	17 0	6 **−3** *−18*	2 **−6** *−12*	1 **−9** *−9*							
22.5	414	−4	−1656	6624	2 **20** *40*	173 **16** *2768*	185 **12** *2220*	41 **8** *328*	9 **4** *36*	3 0	1 **−4** *−4*									
17.5	23	−5	−115	575	2 **25** *50*	16 **20** *320*	4 **15** *60*	1 **10** *10*												
	5317		**−6684**	**35149**																

the items of columns ($\dot{C}$) and (D). These fd^2 products are used to compute σ_X and σ_Y.

The student is already familiar with this method of computing the mean and standard deviation by the short method. The results from the table are:

(1) For the distribution of ages of all husbands (X).

$$c_X = \frac{\Sigma fd}{N} = \frac{-5471 + 5805}{5317} = \frac{+334}{5317} = +.0628 \text{ intervals,}$$

or, $c_X = +.0628$ intervals times 5 years $= +.314$ years,

and Mean $= G.A. + c_X = 42.5$ years $+ .314$ years $= 42.8$ years.

$$\text{Also } \sigma_X{}^2 = \frac{\Sigma fd^2}{N} - c_X{}^2 = \frac{36518}{5317} - (+.0628)^2 = 6.8629,$$

and $\sigma_X = \sqrt{6.8629} = 2.62$ intervals $= 13.1$ years.

(2) For the distribution of ages of all wives (Y),

$$c_Y = \frac{-6684 + 4619}{5317} = \frac{-2065}{5317} = -.3884 \text{ intervals,}$$

or, $c_Y = -.3884$ intervals times 5 years $= -1.942$ years,

and Mean $= 42.5$ years $- 1.942$ years $= 40.6$ years.

$$\text{Also } \sigma_Y{}^2 = \frac{35149}{5317} - (-.3884)^2 = 6.4598,$$

and $\sigma_Y = \sqrt{6.4598} = 2.54$ intervals, or 12.7 years.

Summarizing these results *in intervals:*

(1) $c_X = +.0628$ intervals; $\sigma_X = 2.62$ intervals
(2) $c_Y = -.3884$ intervals; $\sigma_Y = 2.54$ intervals

The product-deviations (d_Xd_Y) from the table. The table is divided into four quadrants by the column and the row in which the $G.A.$ is located and for which the deviations are zero. The signs of the deviations (d_Xd_Y) from the $G.A.$ for any compartment of a given quadrant are shown in the diagram on page 273.

The product-deviation method of obtaining r, instead of reducing the arrays of the columns and rows to means, gives each value in the table as originally tabulated an influence depending upon the amount of its deviation from the mean of X and the mean of Y. Each pair of deviations (d_Xd_Y) may be obtained for any desired compartment from row (3) and column (C), with the proper signs. For example, the lower left-hand compartment of the table enclosed in heavy rulings is located opposite −5 in row (3) and −5 in column (*C*), 5 intervals below the mean of *X* and the same amount below the mean of *Y*. The product-deviation is −5 times −5 = 25. This product is further multiplied by the frequency 2, which gives 50, the *weighted product-deviation.* All frequencies are entered at the top of each compartment in light-face type; all product-deviations appear in heavy-face type; and all weighted product-deviations are entered below the other values in each compartment in italics.

II $d_X = -$ $d_Y = +$ Product = −	I $d_X = +$ $d_Y = +$ Product = +
III $d_X = -$ $d_Y = -$ Product = +	IV $d_X = +$ $d_Y = -$ Product = −

In quadrants I and III the associated deviations (d_Xd_Y) are either both above their respective means or both below, and the resulting products are positive; in quadrants II and IV a deviation below the mean is associated with a deviation above the mean and the products are negative. In other words, *like deviations* are found in I and III, and *unlike deviations* in II and IV. In this table the like deviations and the positive products greatly predominate and the *correlation is positive.* If the frequencies are massed along a diagonal from the left-upper to the right-lower corner of the table, the unlike deviations and the negative products will predominate in quadrants II and IV and the *correlation will be negative.*

An algebraic sum (Σd_Xd_Y) *of all the weighted product-deviations shows to what extent the like deviations or positive products predominate in the table.* This is obtained by adding all figures in italics for quadrants I and III ($+d_Xd_Y$) and subtracting from the sum the total of the figures in italics in quadrants II and IV ($-d_Xd_Y$), which equals 32,244 (the algebraic sum of d_Xd_Y).

But this is a summation of the product-deviations from the two *assumed means*, not from the *true means.* Since the assumed means are in error by the amounts c_X and c_Y, each deviation, d_X and d_Y, is in error by the amount of the correction. We must apply a correction to obtain Σxy, the *sum of the product-deviations about the true means.*

$$\text{Now, } r = \frac{\frac{\Sigma d_X d_Y}{N} - c_X c_Y}{\sigma_X \sigma_Y} = \frac{\frac{32244}{5317} - (+.0628 \text{ times } -.3884)}{2.62 \text{ times } 2.54}$$

$$= \frac{6.0643 - (-.0244)}{6.6548} = \frac{6.0643 + .0244}{6.6548}$$

$$= +.91$$

All the computations are in intervals. The value of r (+.91), computed by this more exact method, is almost identical with that obtained from Figure 32.

THE UNRELIABILITY OF THE COEFFICIENT OF CORRELATION

The mere statement of the value of r computed from a single random sample is not necessarily conclusive evidence of the degree of association. The same question discussed at length in Chapter XI arises in this connection. How much fluctuation in the value of r may be expected, due to the *chance conditions of sampling?* If a very large number of similar random samples of the same population are examined and related, *r is shown to be itself a variable*, and the obtained values of r, when arranged in a frequency distribution, tend to assume the bell-shaped form.[1] How much may the r obtained from a second random sample be expected to differ from the r computed from the sample under consideration? The significance of r evidently depends upon the amount of this probable variation which is due to the uncontrolled conditions of sampling, and *the size of r must be judged in relation to the amount of its probable error.* The formula for measuring the probable variation in r is,

$$P.E._{\cdot r} = .6745 \; \frac{1 - r^2}{\sqrt{N}}.$$

This means that the chances are even (1 to 1) that the value of r obtained from another random sample will be located within the range $\pm P.E.$ from the computed r. In this manner $P.E.$ measures the unreliability of a computed coefficient of correlation.

For the correlation between the ages of husbands and wives,

$$P.E._{\cdot r} = .6745 \, \frac{1 - (+.91)^2}{\sqrt{5317}} = \pm .0016$$

and r should be written +.91 ±.0016. The chances are even (1 to 1) that the r from any similar random sample will be as likely as not to fall

[1] The reader should review the discussion of probable error in Chapter XI.

within .0016 plus or minus from .91. In other words, 50 per cent of the values for r, obtained from an indefinitely large number of similar samples, would be located between .9084 and .9116. Likewise, the chances are 142 to 1 (practical certainty) that the value of r obtained from another random sample will not differ more than 4 $P.E.$ from the computed r (.91 ± .0064). In this manner limits for the chance fluctuations of r can be established and its *unreliability* can be determined from a knowledge of probable error related to the size of r.

Significance of a coefficient. The coefficient +.91 is high and the probable error .0016 is small, which indicates that only to a slight extent is the value of r in this problem subject to *variations of a chance character due to sampling. Conservative statistical practice in interpreting r requires that the size of the coefficient should be* 4 *P.E. before it becomes indicative of any significant degree of association.* In this example r is many times its $P.E.$

Experimenting with the formula shows that the size of $P.E.$ and consequently the degree of unreliability increases as the number of cases (N) decreases. Likewise, r increases in reliability as $P.E.$ decreases and as N increases. The reliability does not increase directly as N increases, but in proportion to the square root of N. To cut down the unreliability, one half would require four times the number of cases.

Suppose $r = +.25$ and $N = 36$. Then

$$P.E._{\cdot r} = .6745 \frac{1 - (+.25)^2}{\sqrt{36}} = \pm.11, \text{ and } r \text{ should be stated } +.25 \pm.11.$$

The coefficient without qualification indicates a low degree of association, and the $P.E.$ is so large relative to the size of r, on account of the small number of cases, that a variation of minus 2 $P.E.$ would reduce the value of r to almost zero (.25 ± .22). A variation of 3 $P.E.$ (±.33) might reverse the sign of r, making the coefficient negative. It is evident that a coefficient subject to such reversal on account of the accidental conditions of sampling is not significant of actual relationship.

But suppose $r = +.25$ and $N = 900$. Then

$$P.E._{\cdot r} = .6745 \frac{1 - (+.25)^2}{\sqrt{900}} = \pm.02, \text{ and } r = +.25 \pm.02.$$

The same size of r in this case, when obtained from a much larger number of cases, has a greater degree of reliability and indicates a low degree of association. Even though affected by chance fluctuations the *coefficient is significant.*

It is clear that if the $P.E._{\cdot r}$ is neglected, a high coefficient obtained from the association of few cases may be given the same significance in interpretation as a high coefficient obtained from many items, whereas the one r is much more subject to the variations due to chance conditions of sampling than the other. Likewise, a coefficient of moderate size (.4), computed from a small number of items, may be interpreted as indicating a moderate degree of association, when really its large $P.E.$ makes such an interpretation of doubtful reliability. *It is strongly emphasized that an r of low value to be significant must be based upon many more cases than one of high value.*

By the use of the formula it is easy to construct a table of values for $P.E._{\cdot r}$, computed for different numbers of items and r's of different amounts. It should be noted that where only a small number of cases are available for correlation, unless r is close to unity, the $P.E.$ has little or no value as a measure of significance, because it is so large relative to the size of the coefficient.

THE PREDICTION EQUATIONS FOR Y AND X

Instead of fitting the *straight lines of generalized relationship* to the means of the columns and rows by *inspection*, as was done in Figures 32 and 33, we can now use the equations for which the constant values r, σ_X and σ_Y have been computed directly from the correlation table. By substituting known values in these equations, points on the lines which we wish to draw may be located and the lines may be drawn through these points.

Let X = age of husband, and $\overline{X}$ = mean age, 42.8 years
Y = age of wife, and $\overline{Y}$ = mean age, 40.6 years
Let x = deviation of age of husband from the mean
y = deviation of age of wife from the mean
$\sigma_X = 13.1$ years, and $\sigma_Y = 12.7$ years

The straight line fitting the means of the columns is described by the equation,

$$y = r\,\frac{\sigma_Y}{\sigma_X}\,x, \text{ or } y = +.91\left(\frac{12.7}{13.1}\right)x, \text{ or}$$

(1) $y = .88\,x$ (instead of $y = .87\,x$, obtained by the inspection method)
Another form of the equation describing the line is stated

$$Y - \overline{Y} = r\frac{\sigma_Y}{\sigma_X}(X - \overline{X}), \text{ or } Y - 40.6 \text{ years} = .88\,(X - 42.8 \text{ years})$$

or, transposing, $\quad Y = .88\,X - 37.7 + 40.6,$

(2) $\quad Y = .88\,X + 2.9$ years

Equation (1) means that for every unit variation in the age of husbands we may expect *on the average* about .88 as much variation in the age of wives. In equation (2) we can substitute any specific values of X (age of husbands) and compute the most probable value for the corresponding Y (age of wives). The slope of the line of relationship is determined by $r\frac{\sigma_Y}{\sigma_X} = .88$, which is sometimes called *the coefficient of regression of Y on X*. These equations describe the line of relationship between the actual ages of husbands and the mean ages of their wives.

Let us refer to Figure 32 and fit the line RR_1 to the dots by determining values from the equation $Y = .88\,X + 2.9$, and locating points on the line. Substituting the value of X at A on the diagram (70 years) we have,

(*a*) $Y = .88(70) + 2.9 = 64.5$ years, the ordinate of the point B (by *inspection* the line was drawn through a point the ordinate of which was 64.3 years)

(*b*) Substituting the value of X at C on the diagram (80 years) we have, $Y = .88\ (80) + 2.9 = 73.3$ years, the ordinate of the point D (by *inspection* the line was drawn through the point, $Y = 73.0$ years)

(*c*) Substituting the mean of X (42.8 years) we have, $Y = .88\ (42.8) + 2.9 = 40.6$ years, the mean of Y.

We can now locate three points on the required line of relationship of Y on X, by knowing their coördinate X's and Y's. Any number of other points could be located on it in a similar manner. In order to draw the line RR_1 it is sufficient to locate one other point besides the center of the system. The mean age of the wives (40.6 years) was obtained by substituting the mean age of husbands in the equation so as to verify the assumption that the regression line of Y on X must pass through the center of the system of related values at 0_1. The line RR_1 in the diagram, which was fitted by inspection, should be slightly shifted to conform to the more exact location as determined by the equation, $y = .88x$, instead of $y = .87x$. (See Figure 32, page 260.)

The straight line fitting the means of the rows is described by the equation,

$$x = r\frac{\sigma_X}{\sigma_Y}\,y, \text{ or } x = +.91\left(\frac{13.1}{12.7}\right) y, \text{ or}$$

(1) $x = .94y$ (instead of $x = .93y$, obtained by the inspection method)
Another form of the equation describing the line is stated,

$$X - \overline{X} = r\,\frac{\sigma_X}{\sigma_Y}\,(Y - \overline{Y}\,), \text{ or } X - 42.8 \text{ years} = .94\ (\text{Y} - 40.6 \text{ years}),$$

or, transposing, $X = .94\,Y - 38.2 + 42.8$

(2) $X = .94\,Y + 4.6$ years

Equation (1) means that for every unit change in the age of wives we may expect on the average about .94 as much variation in the age of their husbands. By substituting specific values of Y (ages of wives) in equation (2) we can compute values for the corresponding X's (ages of husbands), which will enable us to locate points on the line of relationship between the actual ages of wives and the mean ages of their husbands. This is the regression line of X on Y, the slope of which is determined by $r \frac{\sigma_X}{\sigma_Y} = .94$. In Figure 33 the line CC_1 was fitted by inspection and its slope was .93. The student should locate this line accurately by the use of the equation $X = .94\,Y + 4.6$ years, in the same manner as explained for RR_1 in Figure 32. The coördinates of points on the line CC_1 may be obtained by substituting values of Y in the equation and computing the corresponding values for X. (Figure 33, page 267.)

The two regression equations furnish a more complete description of relationship than the coefficient of correlation alone. The latter is a pure number indexing the degree of association between two variables. It does not enable us to state how much variation on the average may be expected in one trait when we know the amount of variation in another related series. But the equations of the lines of relationship describing the law of association make possible the prediction of most probable values of one trait from known values of another. The coefficient of correlation is an intermediate measure in describing the law of association. It is the geometric mean of the regression coefficients (.88 and .94 in the correlation of the ages of husbands and wives). It summarizes in one expression the relationships indicated by the regression equations.

Limits of error in the use of the prediction equations.[1] We have used in the preceding section two equations for estimating most probable values in one series from known values in the other series. By their use we are able to locate the straight lines of most probable relationship.

$$Y = .88\,X + 2.9 \text{ years}$$
$$X = .94\,Y + 4.6 \text{ years}$$

The values for Y corresponding to values for X fall upon the straight line RR_1 in Figure 32. But this line is a *generalized experience* and not the actual facts of the columns of the correlation table. The ages of wives (Y) are *actually scattered in a sub-distribution* in each column. *Therefore, a predicted single value for Y corresponding to a given value for X is subject to a range of error determined by the scatter about the straight line.* (See Figure 34 in which RR_1 is the same as in Figure 32.) The less

[1] In using the regression equations for purposes of prediction, the assumption is made that the distributions of cases about the lines of relationship are bell-shaped.

this scatter, the more closely we approach the condition where there would be only one value for Y for a given value of X. The following formula describes the amount of this scatter (S) in the Y variable about the line of most probable relationship, (RR_1 in Figure 34), in the same sense that the standard deviation describes it for a single distribution.

$$S_Y = \sigma_Y \sqrt{1 - r^2} = 12.7 \sqrt{1 - (+.91)^2} = 5.2 \text{ years}$$

On the hypothesis that the distribution of cases approximates the bell-shaped form, this means that *in about 68 per cent of the cases the actual values of Y will not differ more than ± 5.2 years from the predicted values.* In other words, if many predictions of the ages of wives are made from known ages of husbands, the chances are 2 to 1 that the ob-

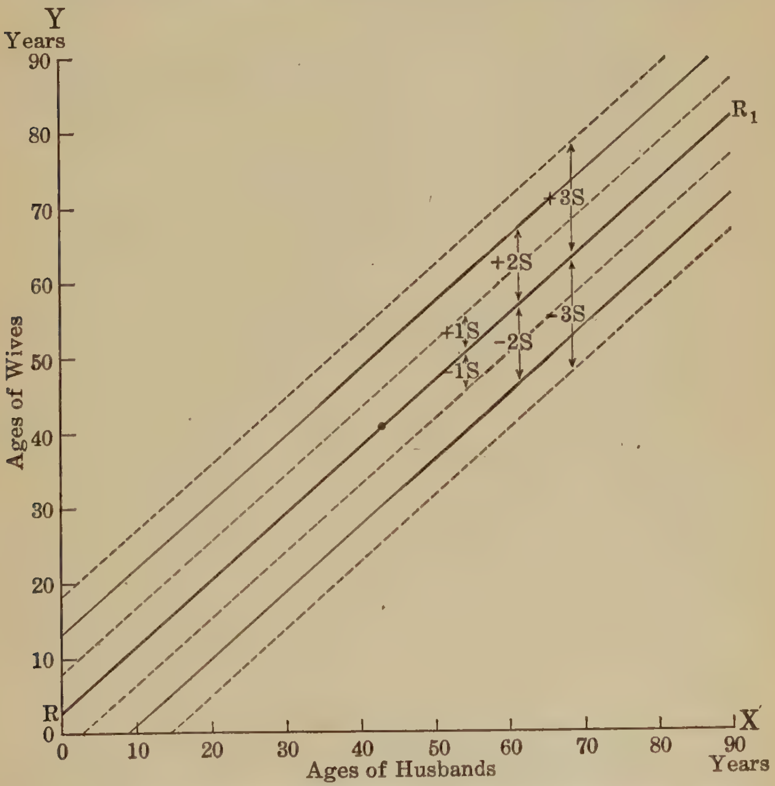


FIG. 34. DISTRIBUTION OF THE ERRORS OF ESTIMATE ABOUT THE LINE OF AVERAGE RELATIONSHIP, RR_1

($S_y = \pm$ 5.2 years, for the ages of wives estimated from known ages of husbands by the use of the regression equation of Y on X, $Y = .88X + 2.9$ years.)

served ages of wives will be distributed within a zone ± 5.2 years from the predicted or most probable ages, which fall upon the line of average relationship. Therefore, about two thirds of the predicted ages may be expected to fall within 5.2 years, plus or minus, of the actual ages. In this manner *limits are defined within which our predictions are likely to be true* according to specified chances ($\pm 1S$, chances 2 to 1; $\pm 2S$, chances 21 to 1; $\pm 3S$, chances 369 to 1, practical certainty). Any predicted value of Y should be written ± 5.2 years, which describes the *probable exactness of the prediction.* (See Figure 34 for representation of the distribution of *errors of estimate* about line RR_1 by lines parallel to it.)

The measure of scatter about the mean of any column of the table is the standard deviation of that column, and within this range plus and minus from the mean of the column about two thirds of the cases are included. It will be remembered that these means of the columns are the values from which the dots are plotted in Figure 32. The means fall closely about the line of most probable relationship RR_1. Therefore, the line connecting the means almost coincides with RR_1. If our hypothesis of linear relationship is true, the means of the columns diverge from the straight line on account of limitations in the sample. The different parts of the entire distribution represented by the sub-distributions in the columns are not entirely homogeneous in their arrangement, especially at the margins of the table where there are few cases.

If our sample could be indefinitely increased until we have an ideal distribution, the means of the columns would fall exactly on the line of most probable relationship, which represents idealized experience. In this case if we should take the sum of the weighted squared deviations in each column about the mean of that column (Σfx^2), combine these sums for all the columns, divide by the total items (N) in the entire distribution, and extract the square root, we would obtain *a measure of scatter for the entire distribution* exactly equal to S_Y in the formula. (See Table 52.)

Observing that the means of the columns in Figure 32 do not fall exactly on the straight line RR_1, we shall follow the procedure just suggested for measuring scatter and ascertain how much the result differs from that obtained from the formula, ($S_Y = 5.2$ years).

Summing column (3), Table 52, dividing by the total number of items in the correlation table and extracting the root, we have

$$\sqrt{\frac{139{,}819.06}{5317}} = \sqrt{26.30} = 5.1 \text{ years,}$$

which is a measure of scatter of the entire number of items *about the means of the columns,* not about the straight line of regression of Y on X.

TABLE 52. SCATTER ABOUT THE MEANS OF THE COLUMNS[a]

(Mean ages of wives corresponding to actual ages of husbands)

X AGES OF HUSBANDS (years) (1)	Y MEANS OF WIVES' AGES (years) (2)	Σfx^2 (for each column) SUMS OF WEIGHTED SQUARED DEVIATIONS FROM THE MEANS OF THE RESPECTIVE COLUMNS (3)
15–19	20.0	25.00
20–24	23.4	2,000.40
25–29	26.9	8,748.68
30–34	31.1	13,453.32
35–39	35.6	16,374.73
40–44	40.2	17,551.00
45–49	44.9	17,862.20
50–54	49.5	16,302.00
55–59	54.0	14,530.25
60–64	58.4	12,090.37
65–69	62.6	9,421.75
70–74	66.4	6,183.84
75–79	69.3	3,738.00
80–84	73.3	1,412.52
85–	75.0	125.00
		139,819.06

[a] The procedure in this table is equivalent to obtaining the standard deviation of each column of the correlation table separately and combining these into a single measure by first squaring the standard deviations and then weighting the squares by the number of items in the respective columns, summing the products, dividing by the total items and taking the square root.

Each quantity in column (3) is the sum of the weighted squared deviations (Σfx^2) about the corresponding mean given in column (2). The reader can verify the items of columns (1) and (2) from Table 50.

It will be noted that the result 5.1 years is *almost identical* with that obtained by the scatter formula (S_Y = 5.2 years), because the means fall so near the straight line. If the line of the means were identical with the straight line of regression the results obtained by the two methods of measuring the scatter would be identical.

Exactly the same procedure can be followed in measuring the scatter about the *means of the rows* of the correlation table. The formula for the computation of the scatter about the *straight line of regression* of X on Y, CC_1 in Figure 33, is

$$S_X = \sigma_X \sqrt{1 - r^2} = 13.1 \sqrt{1 - (+.91)^2} = 5.4 \text{ years.}$$

Any value of X predicted from a known value of Y should be written ± 5.4 years, which means that in about two thirds of the cases the predicted value of X will not differ from the observed values by more than ± 5.4 years.

TABLE 53. CORRELATION BETWEEN SIZE OF FAMILY INCOME AND THE PERCENTAGE SPENT FOR FOOD [a]

	(A)	(B)	(C)	(D)	(E)	SIZE OF ANNUAL FAMILY INCOME (X)																		
(1)	m					$550	650	750	850	950	1050	1150	1250	1350	1450	1550	1650	1750	1850	1950	2050	2150	2250	
(2)		f				4	12	24	25	18	13	13	21	13	22	17	8	4	2	2	1	0	1	**200**
(3)			$d \rightarrow X$			−7	−6	−5	−4	−3	−2	−1	0	1	2	3	4	5	6	7	8	9	10	
(4)			$\downarrow$ Y	fd		−28	−72	−120	−100	−54	−26	−13	**−413**	13	44	51	32	20	12	14	.8	0	10	**+204**
(5)	%				fd^2	196	432	600	400	162	52	13		13	88	153	128	100	72	98	64	0	100	**2671**
PERCENTAGE SPENT FOR FOOD (Y)	57	2	8	16	128		1 **−48**	1 **−40**																
	55	4	7	28	196			2 **−35**	2 **−28**															
	53	3	6	18	108		1 **−36**	1 **−30**				1 **−6**												
	51	13	5	65	325		2 **−30**	3 **−25**	2 **−20**	2 **−15**	1 **−10**		1 0	1 **5**					1 **30**					
	49	10	4	40	160	1 **−28**	1 **−24**	4 **−20**	1 **−16**	2 **−12**			1 0											
	47	18	3	54	162	2 **−21**	1 **−18**	5 **−15**	3 **−12**	1 **−9**		1 **−3**		3 **3**	1 **6**			1 **15**						
	45	14	2	28	56		1 **−12**	3 **−10**	1 **−8**	2 **−6**	1 **−4**		1 0		2 **4**	1 **6**	2 **8**							
	43	16	1	16	16	1 **−7**	1 **−6**	1 **−5**	3 **−4**	3 **−3**		1 **−1**	3 0	1 **1**	1 **2**	1 **3**								

Percentage Spent for Food (*Y*)

41	26	0	+265			1 0		5 0	1 0	2 0	5 0	4 0	3 0	3 0	1 0		1 0						
39	22	− 1	− 22	22		1 6	1 5	2 4	4 3	3 2	2 1	2 0	1 −1	3 − 2	3 − 3								
37	13	− 2	− 26	52		1 12		2 8	1 6	3 4		4 0		2 − 4									
35	13	− 3	− 39	117			2 15	2 12		1 6		1 0		3 − 6	3 − 9	1 −12							
33	17	− 4	− 68	272				1 16	1 12		2 4	2 0	4 −4	1 − 8	2 −12	2 −16	1 −20	1 −24					
31	7	− 5	− 35	175		1 30				1 10				2 −10	2 −15	1 −20							
29	8	− 6	− 48	288					1 18	1 12	1 6	1 0		2 −12	1 −18							1 −60	
27	5	− 7	− 35	245				1 28				1 0			2 −21				1 −49				
25	4	− 8	− 32	256												1 −32	1 −40		1 −56	1 −64			
23	4	− 9	− 36	324			1 45							1 −18	1 −27	1 −36							
21	1	−10	− 10	100										1 −20									
	200		**−351**	**3002**																			

[a] "Analysis of the Standard of Living in the District of Columbia in 1916," by W. F. Ogburn, *Quarterly Publication of the American Statistical Association,* June, 1919, Table 1, p. 377.

Scale of frequency table

NEGATIVE OR INVERSE CORRELATION

Association is described as negative when variation in one series is accompanied by variation in the other, but in the opposite direction. For example, as the size of the annual income of families increases, the percentage of their total annual expenditures devoted to food decreases. If the degree of association (r) between these two series of facts for a large number of families in a community is known, and if the law of their association can be described by the straight line, it will be possible from a knowledge of the size of the annual income to estimate the most probable percentage of the total expenditures which is spent on the average for food. Then the actual expenditure may be compared with the estimated amount which is the generalized experience of the group.

In Table 53 the incomes of 200 families are classified in \$100 intervals on the horizontal scale (X), the rows of the table, and each of the income sub-groups is again classified according to the percentage spent for food in 2-per-cent intervals on the vertical scale (Y), the columns of the table.

Rows (1)–(5) and columns (A)–(E) furnish the data for both series necessary for the computation of the means, c_X, c_Y, σ_X, σ_Y, in the same manner as in Table 51. In the separate compartments of the table the figures in light-face type are the frequencies and those in heavy-face type are the product-deviations in intervals from the guessed average of each series. There is one difference from Table 51. *The weighted product-deviations are not entered.* Therefore, before an algebraic sum of the d_X d_Y products is taken, each figure in heavy type must be multiplied by the corresponding frequency in the given compartment. The weighted products in quadrants II and IV are then added and from this total is subtracted the sum of the weighted products in quadrants I and III. *The negative products of II and IV predominate in this case and the correlation is negative or inverse.*[1]

[1] Various short methods of computing and summing the weighted product-deviations from the correlation table have been devised. Very properly these place the emphasis upon *technique*. The author recognizes the desirability of saving time and labor and the utility of some of these methods for that purpose, but he regards it of first importance to keep the attention of the elementary student upon the *logical significance of the correlation procedure*. For illustrations of short methods the reader may refer to Pearl: *Medical Biometry and Statistics*, chap. XIV, Table 71, and explanation, pp. 305–07; Yule: *Introduction to the Theory of Statistics* (6th ed., 1922), pp. 182–88; Toops: "Eliminating the Pitfalls in Solving Correlation: A Printed Correlation Form," *Journal of Experimental Psychology*, vol. IV, no. 6, December, 1921.

The computations from Table 53 are:

(1) *For the X series, the size of incomes,*

$$c_X = \frac{-413 + 204}{200} = -1.045 \text{ intervals times } \$100 = -\$104.50$$

Then, the mean income = $1250 − $104.50 = $1145.50.

$$\text{Also } \sigma_X = \sqrt{\frac{2671}{200} - (-1.045)^2} = 3.50 \text{ intervals} = \$350.$$

(2) *For the Y series, the percentage spent for food,*

$$c_Y = \frac{-351 + 265}{200} = -.43 \text{ intervals times 2 per cent}$$
$$= .86 \text{ per cent.}$$

Then, the mean percentage = 41.00 per cent −.86 per cent
= 40.14 per cent.

$$\text{Also } \sigma_Y = \sqrt{\frac{3002}{200} - (-.43)^2} = 3.85 \text{ intervals times 2 per cent}$$
$$= 7.70 \text{ per cent.}$$

Summarizing the results in intervals,

$$\begin{aligned} c_X &= -1.045 \text{ intervals;} \qquad & \sigma_X &= 3.50 \text{ intervals} \\ c_Y &= -\ \ .43 \text{ intervals;} \qquad & \sigma_Y &= 3.85 \text{ intervals} \end{aligned}$$

(3) The coefficient of correlation,

$$r = \frac{\dfrac{-1292}{200} - (-1.045 \text{ times } -.43)}{3.50 \text{ times } 3.85} = -.513$$

The quantity −1292 is the algebraic sum of the weighted product-deviations obtained from the four quadrants of the table, and *measures the dominance of the negative values in quadrants II and IV.* These negative values result from the association of variations below the mean in the one series with variations above the mean in the other series and *vice versa,* the product of a plus deviation by a minus deviation producing a negative product-deviation in either case.

(4) $$P.E._{\cdot r} = .6745 \frac{1 - (-.513)^2}{\sqrt{200}} = \pm .035$$

Therefore, $r = -.513 \pm .035$

This coefficient of correlation is not only negative but is located at about the mid-point on the scale from zero to unity, in contrast to the +.91 for the ages of husbands and wives. *Unity would signify absolute dependence* of one variable on the other, which means that for each value of X there would be only one rigidly defined value for Y, instead of the scatter in the sub-distributions of the columns. Therefore, $-.513$ indicates a moderate degree of association, and since the probable error is low, the fluctuation in the value of r to be expected from accidental conditions of sampling is relatively slight.

The regression equations of Y on X. The required constants computed from the correlation table are:

$\bar{X}$ (mean income) = \$1145.50; $\bar{Y}$ (food expenditure) = 40.14 per cent

$$\sigma_X = \$350 \qquad\qquad \sigma_Y = 7.7 \text{ per cent}$$
$$r = -.513$$

Then,

$$(1) \qquad y = -.513\left(\frac{7.7}{350}\right)x = -.011x,$$

in which $-.011$ describes the slope of the regression line of Y on X. This means that for each unit variation in the size of family income, *on the average* about .011 as much variation may be expected in the percentage spent for food, but *in the opposite direction.* In other words, when the income increases the proportion of total expenditures devoted to food declines.

The equation needed for predicting the most probable values of Y corresponding to specific values of X is stated:

$$Y - 40.14 \text{ per cent} = -.011\ (X - \$1145.50) \text{ or, transposing}$$
$$(2) \qquad Y = -.011\,X + 52.74$$

which describes the *straight line of generalized relationship of Y on X.*

Locating the regression line of Y on X. By substituting values of X in equation (2), corresponding most probable values for Y are easily obtained. The related values for X and Y are the coördinates of points on the line of regression.

First, let X = \$850

then $Y = -.011\ (\$850) + 52.74 = 43.39$ per cent.

If we calculate the actual mean percentage of the column in the correlation table which corresponds to \$850 on the horizontal scale, the result is 42.52 per cent, showing that the mean of that column is close to the line of most probable relationship.

Now, let $X = \$1250$

then $Y = -.011\ (\$1250) + 52.74 = 39.0$ per cent.

The actual mean percentage of the column opposite $1250 on the horizontal scale is 39.1 per cent, almost identical with the predicted value (39.0), which falls on the line of most probable relationship.

Having the coördinates X and Y for two points on the regression line ($X = \$850$, $Y = 43.39$ per cent; and $X = \$1250$, $Y = 39.0$ per cent) at C and D, it is a simple matter to draw a straight line RR_1 through these points, in Figure 35. This line passes through the center of the system of related values as zero, shown by substituting the mean income, $1145.50, in the equation. We have,

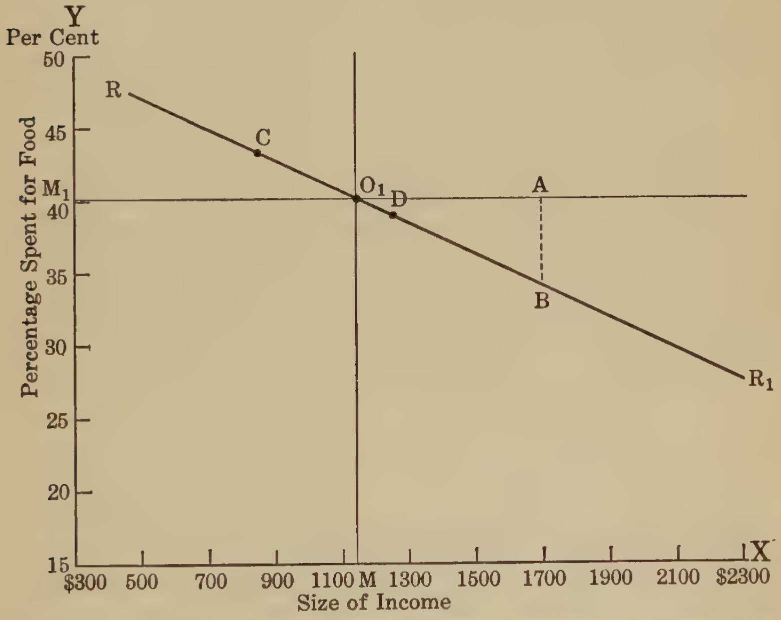


FIG. 35. RELATIONSHIP BETWEEN SIZE OF ANNUAL INCOME AND PERCENTAGE OF THE TOTAL SPENT FOR FOOD

The line of average relationship of Y on X, RR_1. (Equation of straight line RR_1 is $y = -.011\ x$, with origin at O_1. If the origin is taken at zero, then $Y = -.011X + 52.74$.)

$$Y = -.011\ (\$1145.50) + 52.74 = 40.14 \text{ per cent,}$$

the mean percentage spent for food.

The line RR_1 in Figure 35 is the regression line of Y on X. Since the association is *negative* or *inverse*, the direction of the line is downward across the page from left to right. As the income increases the percentage spent for food, on the average, declines. The slope of the line is determined by the regression coefficient

$$r\frac{\sigma_Y}{\sigma_X} = -.011$$

which means that a change in the size of income (O_1A) is associated with .011 as much variation in the percentage for food, ($A\ B$).

Measure of scatter about the regression line of Y on X. The probable exactness of the predicted values of Y corresponding to known values of X is determined by the scatter about the line RR_1. This scatter is described by the formula

$$S_Y = \sigma_Y\sqrt{1 - r^2} = 7.7\sqrt{1 - (-.513)^2} = 6.6 \text{ per cent.}$$

Any estimated values of Y obtained by the use of the equation should be written $\pm S_Y$, as 43.39 ± 6.6 per cent, and 39.0 ± 6.6 per cent. In this case the scatter is relatively large and the coefficient of correlation (r) is not high. The zone (±6.6) within which two-thirds of the observations may be expected to fall is wide in relation to the line of most probable relationship. This limits the value of this regression equation for purposes of prediction.

CORRELATION OF UNGROUPED DATA

For the correlation of ungrouped data the items are related directly without grouping, and the linear type of association is assumed in the illustration. Table 54 sets forth the relationship between the infant mortality rate, stated in number of deaths under one year per thousand births, and the degree of overcrowding in the housing of the population, as measured by the percentage of the total population of the given district living in private dwellings with more than two persons per room. The data are taken from *London Statistics*, vol. 23, pp. 92–93 and 227, published by the London County Council.

Columns (4) and (5) show the deviations above or below the respective means of the two series. By comparison of these columns it can be seen when deviations with like signs are associated and when the opposite is true. Associated deviations with like signs produce positive product-deviations and those with unlike signs produce negative product-

TABLE 54. CORRELATION BETWEEN OVERCROWDING AND INFANT MORTALITY IN LONDON DISTRICTS IN 1912

DISTRICT	X PER CENT OVER-CROWDED	Y IN-FANT RATE	x DEVI-ATION FROM MEAN	y DEVI-ATION FROM MEAN	x^2	y^2	PRODUCT-DEVIATIONS xy (4) TIMES (5)	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8) +	(9) −
(1) City of London	12.3	81	− 5.6	− 8	31.36	64	44.8	
(2) Battersea	13.3	84	− 4.6	− 5	21.16	25	23.0	
(3) Bermondsey	23.4	111	+ 5.5	+22	30.25	484	121.0	
(4) Bethnal Green	33.2	96	+15.3	+ 7	234.09	49	107.1	
(5) Camberwall	13.5	83	− 4.4	− 6	19.36	36	26.4	
(6) Chelsea	14.9	68	− 3.0	−21	9.00	441	63.0	
(7) Deptford	12.2	89	− 5.7	0	32.49	0	0	
(8) Finsbury	39.8	114	+21.9	+25	479.61	625	547.5	
(9) Fulham	14.6	94	− 3.3	+ 5	10.89	25		16.5
(10) Greenwich	12.1	84	− 5.8	− 5	33.64	25	29.0	
(11) Hackney	12.4	80	− 5.5	− 9	30.25	81	49.5	
(12) Hammersmith	14.2	90	− 3.7	+ 1	13.69	1		3.7
(13) Hampstead	7.1	62	−10.8	−27	116.64	729	291.6	
(14) Holborn	25.6	80	+ 7.7	− 9	59.29	81		69.3
(15) Islington	20.0	87	+ 2.1	− 2	4.41	4		4.2
(16) Kensington	17.1	91	− .8	+ 2	.64	4		1.6
(17) Lambeth	13.6	86	− 4.3	− 3	18.49	9	12.9	
(18) Lewisham	3.9	70	−14.0	−19	196.00	361	266.0	
(19) Paddington	16.2	98	− 1.7	+ 9	2.89	81		15.3
(20) Poplar	20.6	107	+ 2.7	+18	7.29	324	48.6	
(21) St. Marylebone	20.7	93	+ 2.8	+ 4	7.84	16	11.2	
(22) St. Pancras	25.5	88	+ 7.6	− 1	57.76	1		7.6
(23) Shoreditch	36.6	123	+18.7	+34	349.69	1156	635.8	
(24) Southwark	25.8	105	+ 7.9	+16	62.41	256	126.4	
(25) Stepney	35.0	105	+17.1	+16	292.41	256	273.6	
(26) Stoke Newington	8.8	72	− 9.1	−17	82.81	289	154.7	
(27) Wandsworth	6.3	76	−11.6	−13	134.56	169	150.8	
(28) Westminster	12.9	84	− 5.0	− 5	25.00	25	25.0	
(29) Woolwich	6.3	73	−11.6	−16	134.56	256	185.6	
					2498.48	5873	+3193.5	−118.2
Mean[a]	17.9	89						

[a] In column (2) of Table 54 the mean percentage of the population living under conditions of overcrowding (17.9) is obtained by adding the items and dividing by 29. This method of averaging may seem to be a violation of the principles set forth in Chapter VI, where the warning was given that in combining percentages which are computed from aggregates of different sizes, as the populations of these London districts, the percentages should be weighted by the populations to which they apply. This would be true if our primary object in averaging were to obtain a typical percentage of the population living under overcrowded conditions in the entire area.

In correlation, however, the purpose of obtaining an average of percentages for the various districts is different. *In this case the percentages constitute a scale of overcrowdedness from 1 to 100.* Our object is to observe and to measure the association between *variations* in overcrowding and *variations* in the infant death-rate. Therefore, the position of any district on this scale of overcrowdedness is the important consideration, and the mean of all the percentages establishes a typical position for all districts together. We wish to observe whether a district below the mean in overcrowding is also below the mean in infant mortality, and whether a district above the mean in overcrowding is also above in mortality. For this purpose the different sizes of the populations in the various districts are disregarded and a simple average of the percentages is taken. The same reasoning applies to the mean infant death-rate (89) of column (3). The student should remember this illustration in making correlations where percentages or rates are involved.

deviations. The last two columns set forth these plus and minus product-deviations. The standard deviations of X and Y are computed from columns (6) and (7) respectively ($\sigma_X = 9.3$ per cent, $\sigma_Y = 14.2$).

The coefficient of correlation indexing the degree of association between overcrowding and infant death-rates is easily computed by the usual formula: [1]

$$r = \frac{\Sigma xy}{N\sigma_X\sigma_Y} = \frac{3193.5 - 118.2}{29 \text{ times } 9.3 \text{ times } 14.2} = \frac{3075.3}{3829.74} = +.80$$

The probable error is ±.045 and the coefficient is high. This indicates a close relationship between crowding in dwellings and the infant death-rate in London in 1912. To free the infant death-rate from the influences peculiar to the particular year it would have been better to have taken the number of infant deaths per thousand births for three years 1910, 1911, and 1912 combined.

The regression equations for estimating most probable values in the one series from known values in the other series are obtained in the usual manner.

(1) $Y = 1.22\,X + 67.2$ (Regression of Y on X)

(2) $X = .52\,Y - 28.4$ (Regression of X on Y)

NON-LINEAR RELATIONSHIPS BETWEEN TWO VARIABLES

The product-deviation method (r) of measuring the degree of association between two variables is based upon the hypothesis that a straight line fits most closely the means of the columns and the rows in a correlation table, and therefore describes the association in the best possible manner. But sometimes the means conform more closely to some other form of curve. In this case the product-deviation formula and the equa-

[1] A different procedure in computing r is sometimes employed when the standard deviations are not required for the regression equations.

Since $\sigma_X = \sqrt{\frac{\Sigma x^2}{N}}$ and $\sigma_Y = \sqrt{\frac{\Sigma y^2}{N}}$ the formula may be stated:

$$r = \frac{\Sigma xy}{N\sigma_X\sigma_Y} = \frac{\Sigma xy}{N\left(\sqrt{\frac{\Sigma x^2}{N}} \text{ times } \sqrt{\frac{\Sigma y^2}{N}}\right)} = \frac{\Sigma xy}{N\sqrt{\frac{\Sigma x^2 \cdot \Sigma y^2}{N^2}}}$$

$$= \frac{\Sigma xy}{N \text{ times } \frac{1}{N}\sqrt{\Sigma x^2 \cdot \Sigma y^2}} = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \cdot \Sigma y^2}}$$

The denominator of this modified formula does not require the standard deviations but takes the root of the product of the sums of columns (6) and (7) in the table. Of course the value of r is the same whichever procedure is employed.

[illegible] warner

tions of the straight lines (Y on X and X on Y) are not the best possible descriptions of the facts.

When the line of the means is non-linear, the degree of association may be high and yet r will not reveal it. If the linear hypothesis is applied in this situation the coefficient r *always* indexes a lower degree of association than actually exists, because the scatter about the *straight* line is greater than it would be about the *best fitting* line. *A low value for r does not prove that the degree of association is really small or that the two variables are unrelated.*

A more general measure of correlation. It is apparent that we need some general measure of relationship which is applicable to the data whether the lines of the means are linear or non-linear.

In linear correlation we describe the scatter about the straight line of relationship Y on X by the formula, $S_Y = \sigma_Y \sqrt{1 - r^2}$, which defines the limits of error in predicted values of Y. A second formula, $S_X = \sigma_X \sqrt{1 - r^2}$, describes the scatter about the regression line of X on Y, and defines the limits of error in predicted values of X. From either of these formulæ r may be derived. From the first we have:

$$S_Y = \sigma_Y \sqrt{1 - r^2},$$

squaring and transposing, $r^2 = \dfrac{\sigma_Y{}^2 - S_Y{}^2}{\sigma_Y{}^2} = 1 - \dfrac{S_Y{}^2}{\sigma_Y{}^2}$ and

$$(1) \qquad r = \sqrt{1 - \frac{S_Y{}^2}{\sigma_Y{}^2}}$$

Likewise, from the other scatter formula we have:

$$(2) \qquad r = \sqrt{1 - \frac{S_X{}^2}{\sigma_X{}^2}}$$

On the hypothesis that the relationship is linear the coefficient of correlation is stated in formula (1) in terms of a *ratio* of scatter in Y, ($S_Y{}^2$), to the standard deviation of the entire Y distribution ($\sigma_Y{}^2$). In formula (2) r is stated in terms of a *ratio* of scatter in X, ($S_X{}^2$), to the standard deviation of the entire X distribution ($\sigma_X{}^2$). *The scatters (S) are the distributions about the lines of regression and represent the situation where the means of the columns and of the rows fall upon straight lines.*

These formulæ suggest that similar ratios may be used to measure the degree of association when the relationship is non-linear — when the means of the columns and rows are described in best fashion by other forms of curves. In these cases the scatter (represented by S in the formulæ) is no longer measured from a *straight line*. The measure of scatter is the square root of the mean weighted squared deviations of the sub-distributions in the columns or in the rows *about their respective means,* as illustrated in Table 52. This is similar to the concept of the standard deviation in a single distribution. We have already indicated that this measure is identical with S when the means fall exactly on a straight line. The lines of "best fit" are free to take the forms which most closely approximate the means of the columns and of the rows.

If we adopt the linear hypothesis and compute r when the relationship is really non-linear, the contribution of each column or row to S_Y^2 or S_X^2 in formulæ (1) and (2) will be greater than it should be because the scatter (S) is measured about the straight line. This will make the coefficient r less than it should be for a non-linear relationship. Therefore, r may be small because of the assumption of linearity and not because the degree of relationship is really slight. Some other method of measuring association might increase the size of the coefficient and change our interpretation of the results.

For measuring non-linear relationships Pearson has proposed a new constant, η, the *correlation ratio.* Since there are two lines of the means there are also two correlation ratios for each correlation table, computed according to the following procedures.[1]

$$(3) \qquad \eta \text{ (regression of } Y \text{ on } X) = \frac{\sigma \text{ of the means of the columns}}{\sigma \text{ of the entire } Y \text{ distribution}}$$

$$(4) \qquad \eta \text{ (regression of } X \text{ on } Y) = \frac{\sigma \text{ of the means of the rows}}{\sigma \text{ of the entire } X \text{ distribution}}$$

The ratio in (3) may be described further by symbols:

$$(3) \qquad \eta_{YX} = \frac{\sqrt{\dfrac{\Sigma n_X\,(\bar{Y}_X - \bar{Y})^2}{N}}}{\sigma_Y}$$

in which the numerator takes the form $\sqrt{\dfrac{\Sigma fx^2}{N}}$, the usual procedure in

[1] For further explanation of this method and the derivation of formulæ, refer to Yule: *An Introduction to the Theory of Statistics* (6th ed., 1922), pp. 204–07; Kelley: *Statistical Method,* pp. 238–45; and to Pearl: *Medical Biometry and Statistics,* pp. 311–17.

computing a standard deviation, and the symbols have the following significance:

n_X = total frequencies in any column.
$\overline{Y}_X$ = mean of any column of Y's corresponding to given values of X.
$\overline{Y}$ = mean of the entire Y distribution.
N = total number of related items in table.
σ_Y = standard deviation of the entire Y distribution.

The ratio in (4) may be stated:

$$\eta_{XY} = \frac{\sqrt{\frac{\Sigma n_Y\,(\overline{X}_Y - \overline{X})^2}{N}}}{\sigma_X}, \text{ in which} \tag{4}$$

n_Y = total frequencies in any row.
$\overline{X}_Y$ = mean of any row of X's corresponding to given values of Y.
$\overline{X}$ = mean of entire X distribution.
N = total number of related items in table.
σ_X = standard deviation of the entire X distribution.

The lines best fitting the means are free to take any form, either straight lines or other forms of curves. The closer the association the less scatter there is in the columns and the rows, and the more nearly does the standard deviation of the means approach that of the entire distribution. If there were no scatter the standard deviations in the numerator and denominator of the ratio would be identical and η would equal unity. This index of the degree of relationship is a quotient of two standard deviations and, therefore, the sign is *indeterminate.* It follows that η does not reveal whether the association is direct or inverse — whether related variations take place in the same direction from their respective means or in opposite directions.

The student should observe that the correlation ratio is applicable only when the cases are numerous enough to make possible a grouped correlation table. The method is not for use with ungrouped data.

The relation of η to r. *The value of η is either equal to or greater than that of r.* If the relationship is linear the two measures are equal; if non-linear η is larger than r. *The difference between the values of η^2 and r^2 is an index of divergence from linearity of the line fitting the means. Then $\eta^2 - r^2$ is a test of linearity.*[1] This difference is determined separately for the two lines Y on X and X on Y. One line of relationship in a correlation diagram may prove to be linear while the other is non-linear. For any correlation table where linearity is in doubt, both η and r should

[1] Karl Pearson: "On the General Theory of Skew Correlation and Non-Linear Regression," *Drapers' Company Research Memoirs,* Biometric Series II, p. 30 and p. 52.

be computed. *This is especially important when r is low in value*, because the degree of relationship may be concealed by the assumption of linearity. It is not necessary when r is very high as the computation of η would not significantly change the coefficient.[1]

It should be emphasized that, due to fluctuations of sampling, r and η are likely to differ slightly even if the relationship is really linear. Therefore, if the value of η is greater than that of r it does not prove necessarily that the relationship is non-linear. *The apparent difference may not be a significant one.* The observed difference must be compared with the fluctuations which may be expected on account of the sampling process, the standard error (σ) of the observed difference.

The student is already familiar with the probable error of r. An approximate probable error for η is obtained in the same manner, $P.E._{\eta} = .6745 \frac{1 - \eta^2}{\sqrt{N}}$, which indicates that this measure of correlation also is subject to fluctuations due to sampling. Since both r and η are subject to these fluctuations, the difference between them is also a variable. The standard error (σ) of this difference may be roughly stated:

$$\sigma_{(\eta^2 - r^2)} = 2\sqrt{\frac{\eta^2 - r^2}{N}}$$

especially when the difference is small.[2]

The correlation table on page 295 classifies 1000 height and weight measurements. The last column of the table shows the *means of the rows*, the mean height corresponding to given weights; the last row contains the *means of the columns*, the mean weights corresponding to given heights. These data will enable the student to plot the means of Y on X and also the means of X on Y *in order to examine the trend of the means as to linearity.* When the means of the rows (X on Y) are plotted, inspection of the trend leaves doubt as to the linearity of the relationship (Figure 31B, page 255), although the means of the columns (Y on X) appear to conform to this type (Figure 31A, page 254).

We wish to test the linear hypothesis further by computing both r and η for the same data. The student can easily complete by the short method the computation of r from Table 55.

[1] This is illustrated in the correlation of the ages of husbands and wives, where the η test of linearity was not applied.

[2] See Yule: *Introduction to the Theory of Statistics* (6th ed., 1922), p. 352; and Kelley: *Statistical Method*, pp. 238 and 239. For the more exact Blakeman formula, see *Biometrika*, vol. IV, 1905, "On Tests for Linearity of Regression in Frequency Distributions." The formula used is an abridgment of the Blakeman formula.

TABLE 55. CORRELATION BETWEEN HEIGHT AND WEIGHT

			Height (X)																
	m → ↓ Pounds	Inches	60.5	61.5	62.5	63.5	64.5	65.5	66.5	67.5	68.5	69.5	70.5	71.5	72.5	73.5	74.5	75.5		Means of rows (inches) ↓
		f	3	12	21	55	51	135	139	152	154	115	67	45	29	13	8	1	1000	
Weight (Y)	205	1							1											66.5
	195	3							1	1							1			69.5
	185	9							1	2		1	1	3		1				69.9
	175	18				1		3	1	3	4	1	1		1	2	1			68.7
	165	46				1		2	1	3	10	7	9	6	4	3				69.8
	155	89			1	1	1	8	4	9	14	15	16	8	8	2	2			69.3
	145	160			1	3	1	12	15	25	26	31	18	12	9	4	3			68.9
	135	242		1	1	7	5	30	46	42	50	35	8	12	3		1	1		67.8
	125	245			3	17	24	36	37	49	37	23	12	2	4	1				67.0
	115	146	1	5	11	16	15	37	28	17	11	2	2	1						65.6
	105	28		3	3	5	4	6	4	1	2									64.7
	95	13	2	3	1	4	1	1						1						63.3
		1000																		
	Means of Columns (pounds) →		101.7	109.2	118.3	122.6	121.9	128.6	130.9	134.6	137.3	140.6	145.4	146.8	147.8	157.3	156.3	135.0		

Mean height ($\bar{X}$) = 67.6 inches. Mean weight ($\bar{Y}$) = 134.45 pounds

$\sigma_X = 2.6$ intervals $= 2.6$ inches, $\sigma_Y = 1.7$ intervals $= 17.0$ pounds

$c_X = +.072$ intervals, $c_Y = -.055$ intervals

and $\Sigma d_X d_Y = +2633 - 437 = +2196$

$$\text{Therefore, } r = \frac{\dfrac{+2196}{1000} - (+.072 \text{ times} - .055)}{2.6 \text{ times } 1.7}$$

$$= +.498 \pm .016$$

This coefficient indicates a moderate degree of relationship between height and weight. Is this the true degree of relationship?

TABLE 56. THE COMPUTATION OF THE CORRELATION RATIO Y ON X

$\bar{Y}$, the mean of the entire Y series = 134.5 pounds

[a] Means of Columns (pounds) $\bar{Y}_X$ (1)	Deviations from $\bar{Y}$ $(\bar{Y}_X - \bar{Y})$ (2)	$(\bar{Y}_X - \bar{Y})^2$ (3)	Frequency in Each Column n_X (4)	Columns (3) Times (4) $n_X(\bar{Y}_X - \bar{Y})^2$ (5)
101.7	−32.8	1075.84	3	3227.52
109.2	−25.3	640.09	12	7681.08
118.3	−16.2	262.44	21	5511.24
122.6	−11.9	141.61	55	7788.55
121.9	−12.6	158.76	51	8096.76
128.6	− 5.9	34.81	135	4699.35
130.9	− 3.6	12.96	139	1801.44
134.6	.1	.01	152	1.52
137.3	2.8	7.84	154	1207.36
140.6	6.1	37.21	115	4279.15
145.4	10.9	118.81	67	7960.27
146.8	12.3	151.29	45	6808.05
147.8	13.3	176.89	29	5129.81
157.3	22.8	519.84	13	6757.92
156.3	21.8	475.24	8	3801.92
135.0	.5	.25	1	.25
			1000	74,752.19

[a] These means are taken from the last row of Table 55. The value 135.0 represents only one case and is not significant.

$$\eta_{YX} = \frac{\sqrt{\dfrac{\Sigma n_X (\bar{Y}_X - \bar{Y})^2}{N}}}{\sigma_Y}$$

$$= \frac{\sqrt{\dfrac{74{,}752.19}{1000}}}{17.0} = .509$$

We have computed the standard deviation of all the means of the columns about the mean of all the Y's and have related it to the standard deviation of the entire Y distribution. *The result is the correlation ratio of Y on X.*

The coefficient is very little larger than r (+.498.) Could this difference be accounted for by chance fluctuations due to sampling? Let us measure the *unreliability of the difference between* η^2 and r^2, which is the *test of linearity.* Using the abridgment of the Blakeman formula, we have for the standard error (σ):

$$\sigma_{(\eta^2 - r^2)} = 2\sqrt{\frac{\eta^2 - r^2}{N}}$$

$$= 2\sqrt{\frac{(.509)^2 - (.498)^2}{1000}}$$

$$= .0067$$

$$3\,\sigma \text{ of difference } (\eta^2 - r^2) = .02$$

$$\text{Observed difference } (\eta^2 - r^2) = .011$$

It is evident that $3\,\sigma$ of the difference between η^2 and r^2, which would include practically all possible fluctuations to be expected from other similar samples, is about twice the observed difference between η^2 and r^2. Therefore, *the observed difference in the two coefficients is not significant of a real difference*, since it can be accounted for as due to fluctuations of sampling. This indicates that the regression of Y on X is linear. In confirmation of this test of linearity the reader should observe the arrangement of the means in Figure 31A, page 254. The line of the means approximates closely to a straight line except at the extremes where there are few cases. The mean which departs farthest from the trend represents only one height-weight item and cannot be considered as significant. It is not connected with the other means in the diagram for this reason. The reader will observe a contrast between this line of the mean weights and that for the mean heights related to given weights, portrayed in Figure 31B on the opposite page (255). It may happen that one line of the means in a correlation table departs from linearity while the other line of the means is described by a straight line as well as by any other form of curve. We are assuming in this illustration that the 1000 heights and weights are truly representative of the larger population from which they were taken. The contrast between the diagrams indicates that the correlation ratio should be computed also for X on Y. This is presented in Table 57.

TABLE 57. THE COMPUTATION OF THE CORRELATION RATIO X ON Y

$\overline{X}$, the mean of the entire X series = 67.6 inches

[a] MEANS OF ROWS (inches) $\overline{X}_Y$ (1)	DEVIATIONS FROM $\overline{X}$ $(\overline{X}_Y - \overline{X})$ (2)	$(\overline{X}_Y - \overline{X})^2$ (3)	FREQUENCY IN EACH ROW n_Y (4)	COLUMNS (3) TIMES (4) $n_Y(\overline{X}_Y - \overline{X})^2$ (5)
63.3	−4.3	18.49	13	240.37
64.7	−2.9	8.41	28	235.48
65.6	−2.0	4.00	146	584.00
67.0	− .6	.36	245	88.20
67.8	.2	.04	242	9.68
68.9	1.3	1.69	160	270.40
69.3	1.7	2.89	89	257.21
69.8	2.2	4.84	46	222.64
68.7	1.1	1.21	18	21.78
69.9	2.3	5.29	9	47.61
69.5	1.9	3.61	3	10.83
66.5	−1.1	1.21	1	1.21
			1000	1989.41

[a] These means of rows are found in Table 55, in the last column on the right. The value 66.5 represents only one case.

$$\eta_{XY} = \frac{\sqrt{\frac{\Sigma n_Y(\overline{X}_Y - \overline{X})^2}{N}}}{\sigma_X}$$

$$= \frac{\sqrt{\frac{1989.41}{1000}}}{2.6} = .542$$

We have computed the standard deviation of the means of the rows about the mean of all the X's and have related it to the standard deviation of the entire X distribution. *This is the correlation ratio of X on Y.*

What is the *unreliability* of the difference between η^2 and r^2? From the formula we have,

$$\sigma_{(\eta^2 - r^2)} = 2\sqrt{\frac{\eta^2 - r^2}{N}}$$

$$= 2\sqrt{\frac{(.542)^2 - (.498)^2}{1000}}$$

$$= .0136$$

$3\,\sigma$ of difference $(\eta^2 - r^2) = .041$
Observed difference $(\eta^2 - r^2) = .046$

In the regression of X on Y, $3\,\sigma$ of the difference between η^2 and r^2 is slightly less than the observed difference. Since $3\,\sigma$ includes practically all possible fluctuations due to sampling and since the observed difference lies outside this limit, it is highly probable that *the observed difference between the two coefficients is significant of a real difference.* This indicates that the regression of X on Y is probably *slightly non-linear.*[1] In Figure 31B, page 255, the means of the rows are plotted in order that the reader may observe their arrangement in the light of the *test of linearity.*

The correlation ratio is a measure of the degree of association applicable to both linear and non-linear relationships. The r formula and the equations of the straight lines, Y on X and X on Y, apply strictly only to the linear type. But the correlation ratio also has limitations. It is not applicable to ungrouped data. It does not enable us to estimate values in the related series from known values in the given series, as in the case of the regression equations of the straight lines. Its value lies in giving an index of maximum correlation. It furnishes a means of detecting divergence from linear relationship, and its use prevents errors in conclusions due to the wrong assumption.

METHODS OF MEASURING THE DEGREE OF ASSOCIATION FROM THE RELATIVE POSITIONS OF THE VALUES

The methods explained give due importance to the *value and position of each measure in the series.* It has been shown that the arithmetical work involved is rather laborious. Other ways of measuring the degree of relationship have been devised which take account only of relative position in the series of values and not of the exact amount of difference between values. Sometimes it is desired to make a rapid preliminary examination of series of data to test the existence of correlation rather than to measure the exact degree of relationship. Furthermore, the number of pairs of related values which are available may be small. In this case the probable error and the resulting unreliability due to the conditions of sampling have been shown to be large. This unreliability affects the mean, the standard deviation and the measures of relationship by whatever method computed. Therefore, in such cases the methods so far discussed, which are exact and reliable provided the sample is adequate, may secure no more reliable results than simpler and less laborious methods.

The method of correlation from ranks. To meet the needs of research workers interested in psychological and educational data, Professor Spearman worked out *empirically* a method of correlation from ranks or

[1] It should be noted that this slight degree of departure from linearity is usually not regarded, in practice, as substantially affecting the significance of r.

positions in the series. This method has proved very useful in other fields of statistical practice. The coefficient thus obtained is usually designated by ρ (rho), but it should be pointed out that it is identical with r provided the ranks used in the computation constitute the observations or scores of the given trait.

Spearman's formula[1] is usually stated,

$$\rho = 1 - \frac{6\Sigma D^2}{N(N^2 - 1)}$$

in which D is the difference in the corresponding ranks in two related series and N has the usual meaning. This method is based upon the assumption that individual values in a series differ from each other by an equal amount throughout the range of values. This is contrary to the usual arrangement in a frequency distribution, where the items tend to mass about the central value. Mental tests also seem to contradict this assumption and to indicate that abilities are distributed in a form which approximates the bell-shaped distribution. If this be true, then individual scores of mental abilities in the middlé of the range differ from each other much less than do extreme values at either margin of the distribution.

Therefore, unless the ranks constitute the original scores or observations, Spearman's method, in so far as it assumes that ρ is equal to r, proceeds under an assumption contrary to the facts. Usually we have the original scores or observations and then rank them in order, stating the ranks as a new series of values (see Table 58). The ranks are evenly distributed along the scale, but the original observations are not.

On the assumption that the original data from which the ranks have been obtained are arranged in a bell-shaped distribution, Pearson has developed *a correction for the rank formula* by which we can transmute ρ into r. The complete formula for r in terms of ρ is:

$$r = 2 \sin \left(\frac{\pi}{6}\rho\right)$$

The table in Appendix E gives the values of r computed from this for-

[1] See Kelley: *Statistical Method*, pp. 191–94, for derivation of this formula and a statement as to the limitations of Spearman's foot-rule formula, with which this must not be confused. The foot-rule formula is,

$$R = 1 - \frac{6\Sigma G}{N^2 - 1}.$$

The coefficient obtained by the foot-rule method does not vary between -1 and $+1$. It has a large probable error, and has none of the merits, except brevity, of the formula here presented.

mula corresponding to given values for ρ computed from Spearman's rank formula. The correction is of small magnitude.

This is the best rank formula but, of course, there is always some loss in accuracy in changing the original observations into ranks. The probable error of ρ, as determined by Pearson, is about five per cent greater than the probable error of r.

An illustration of the computation of ρ. In Table 58 the number of wage-earners, the cost of materials, the value of products, and the value added by manufacture are given for each of fifteen important industries. In column (2) the industries are ranked according to the number of wage-earners, the industry with the largest number being given first rank. Columns (4), (6) and (8) rank the industries according to the cost of materials, value of products, and value added by manufacture, in the same manner.

TABLE 58. DATA CONCERNING FIFTEEN MANUFACTURING INDUSTRIES, 1919

Industry [a]	Wage-Earners		Cost of Materials		Value of Products		Value Added by Manufacture	
	(1) Number in thousands	(2) Rank	(3) Amount in millions	(4) Rank	(5) Amount in millions	(6) Rank	(7) Amount in millions	(8) Rank
(1) Slaughtering and packing	161	12	$3,783	1	$4,246	1	$ 463	10
(2) Iron and steel	375	5	1,681	3	2,829	2	1,148	2
(3) Automobiles	210	8	1,579	4	2,388	3	809	6
(4) Foundry and machine shop	483	2	948	7	2,289	4	1,341	1
(5) Cotton goods	431	4	1,278	5	2,125	5	847	4
(6) Flour-mill and grist-mill products	45	15	1,799	2	2,052	6	253	15
(7) Petroleum, refining	59	14	1,248	6	1,633	7	385	14
(8) Shipbuilding, steel	344	6	644	12	1,456	8	813	5
(9) Lumber and timber	481	3	471	15	1,387	9	917	3
(10) Cars and shop-construction and railroad repairs	484	1	516	14	1,279	10	763	7
(11) Clothing, women's	166	11	680	10	1,209	11	528	9
(12) Clothing, men's	175	9	606	13	1,163	12	557	8
(13) Boots and shoes	211	7	715	8	1,155	13	440	11
(14) Bread and bakery products	142	13	713	9	1,152	14	439	12
(15) Woolen and worsted goods	167	10	666	11	1,065	15	400	13

[a] *Abstract of the Census of Manufactures, 1919*, p. 19, United States Bureau of the Census.

TABLE 59. CORRELATIONS BY THE RANK METHOD

Industry (from Table 58)	Correlation between Number of Wage-Earners and Value Added by Manufacture					Correlation between Value of Products and Number of Wage-Earners				
	Rank (1)	Rank (2)	D + (3)	D − (4)	D^2 (5)	Rank (6)	Rank (7)	D + (8)	D − (9)	D^2 (10)
(1)	12	10		2	4	1	12	11		121
(2)	5	2		3	9	2	5	3		9
(3)	8	6		2	4	3	8	5		25
(4)	2	1		1	1	4	2		2	4
(5)	4	4	0			5	4		1	1
(6)	15	15	0			6	15	9		81
(7)	14	14	0			7	14	7		49
(8)	6	5		1	1	8	6		2	4
(9)	3	3	0			9	3		6	36
(10)	1	7	6		36	10	1		9	81
(11)	11	9		2	4	11	11	0		
(12)	9	8		1	1	12	9		3	9
(13)	7	11	4		16	13	7		6	36
(14)	13	12		1	1	14	13		1	1
(15)	10	13	3		9	15	10		5	25
			+13	−13	86			+35	−35	482

$$\rho = 1 - \frac{6\,\Sigma\,D^2}{N(N^2 - 1)}$$
$$= 1 - \frac{6\,(86)}{15(225 - 1)}$$
$$= 1 - .15$$
$$= +.85$$

And from Table in Appendix E $r = +.86$

$$\rho = 1 - \frac{6\,\Sigma\,D^2}{N(N^2 - 1)}$$
$$= 1 - \frac{6\,(482)}{15(225 - 1)}$$
$$= 1 - .86$$
$$= +.14$$

And from Table in Appendix E $r = +.15$

In Columns (3) and (4) of Table 59 are given the plus and minus differences (D) between columns (1) and (2). Since the sums of the ranks in columns (1) and (2) are equal, any positive differences in column (3) must be balanced by minus differences in column (4), and the sums of columns (3) and (4) are equal. The same is true of columns (8) and (9). Columns (5) and (10) give the squares of the differences (D^2) between the ranks. It should be noted that the correction for ρ is very slight in transmuting it to r.

The student is urged to correlate the original data, without ranking, in columns (1) and (7) of Table 58, the number of wage-earners and the value added by manufacture. Computing r by the product-deviation method, as was done for ungrouped data on overcrowding and infant death-rates in London districts (Table 54), we find the coefficient to be +.85. *This is almost identical with the value obtained for* ρ *by the Spearman* *formula.* It is suggested that the other correlation be compared in the same manner.

The Census authorities warn the reader that the number of wage-earners and the value added by manufacture are better measures of the relative importance of manufacturing industries than the gross value of products. In some industries the value of materials constitutes by far the larger part of the total value of the finished products, since the labor cost and other expenses are relatively small. Moreover, in some industries there is much more duplication in the gross value than in others, due to the use of the product of one industry or sub-group as materials for another. This duplication does not appear in the value added by manufacture.

The correlations which are shown justify this warning. The degree of association shown by the rank method between the number of wage-earners and the value added by manufacture is high (+.86); whereas, the r for the number of wage-earners related to the gross value of products is very low (+.15). As further evidence the student should compute r by the rank method for the value of products associated with the cost of materials given in Table 58, columns (6) and (4). The value of r is +.73. Thus the degree of association is shown to be fairly high between the cost of materials and the gross value of the products, which is consistent with the statement just made.

By the rank method all computations are simple and the work can be performed very rapidly. It is less accurate than the product-deviation method but proves useful for a rough preliminary examination of the data. It may be used to test the existence of relationship when it may not be adequate to measure the exact degree. The student should note that we cannot use the coefficient obtained by this method for purposes of prediction. In a situation, however, where the number of related items is small and the unreliability is correspondingly great, this method serves as well as a more exact and laborious one.[1]

INTERPRETATION OF THE COEFFICIENT OF CORRELATION

The real difficulties arise in connection with the interpretation of the significance of a given coefficient. Like an average or a measure of dispersion, this measure represents one aspect of the data. It is a pure number ranging between zero and unity, which serves as an index of the degree of association between two series of data. It may assist the student in interpreting the coefficient if we make rather arbitrary subdivi-

[1] For yet cruder methods of measuring relationship, see H. O. Rugg: *Statistical Methods Applied to Education,* pp. 293–99.

sions of this scale from zero to unity, and characterize each in qualitative terms:

(1) A coefficient less than .3, indicates a *low* degree of association and doubtful significance, especially if the number of related items is small.

(2) .3 and less than .5, indicates a *moderate* degree of association if the probable error is small.

(3) .5 and less than .7, indicates *marked* association.

(4) .7 and less than .9, indicates a *high* degree of association.

(5) .9 and over, indicates *very close* association and a very high degree of dependence between the variables.

It is necessary to keep constantly in mind that the interpretation of *significance* is dependent not only upon the size of the coefficient but also upon the number of related items. Especially when the coefficient is small or only of moderate size, the probable fluctuations due to sampling make it unreliable and of doubtful significance if the number of related items is also small. Repeated experiments with many small samples may increase confidence in the results.

The coefficient may be equally consistent with more than one hypothesis. For example, a high positive correlation between overcrowding conditions and infant death-rates may indicate direct dependence of the latter upon the former, or both variables may be dependent upon a third factor, the size of the family income. The coefficient is an index of relationship, not a proof of *causal* dependence. Like other statistical coefficients and constants it is computed for the purpose of clarifying the interpretation of complex masses of data.

The interpretation must be consistent with sound logical analysis. The factors are many and varied in social and economic phenomena. The correlation procedure furnishes a means of testing our hypotheses concerning important factors and their relationships.

Correlation methods as applied to time series will be discussed and illustrated in Chapter XIII.

READINGS

Pearson, Karl, *The Grammar of Science*, chaps. 4 and 5. (Excellent statement of the scientific concept of causation and of the philosophy of correlation.)

Elderton, W. P., and Ethel M., *Primer of Statistics*, chap. 5. (A simple description of the problem of correlation.)

Rugg, H. O., *Statistical Methods Applied to Education*, chap. 9.

Moore, Henry L., *Forecasting the Yield and the Price of Cotton*, chap. 2.

Mills, F. C., *Statistical Methods as Applied to Economics and Business*, chap. 10 ("Linear Correlation"), chap. 12 ("Non-Linear Correlation"), chap. 13 ("The Problem of Estimation"), chap. 14 ("Multiple and Partial Correlation").

Pearl, Raymond, *Medical Biometry and Statistics*, chap. 14. (Tables giving the results of experiments showing the significance of correlation.)

Yule, G. U., *An Introduction to the Theory of Statistics*, 6th ed., chaps. 9 and 10. (Theory and excellent illustrations. Good bibliography of original papers at close of each chapter.)

King, W. I., *Elements of Statistical Method*, chaps. 16 and 17.

Jerome, Harry, *Statistical Method*, chap. 15.

Whipple, G. C., *Vital Statistics*, 2d ed., chap. 14. (Causal relations. Examples in vital statistics.)

Thorndike, E. L., *An Introduction to the Theory of Mental and Social Measurements*, 2d ed., chaps. 10 and 11.

Jones, D. C., *A First Course in Statistics*, chap. 10. (Additional examples in chapter 11.)

Rietz, H. L., and Crathorne, A. R., *Handbook of Mathematical Statistics* (H. L. Rietz, Editor), chap. 8 ("Simple Correlation").

REFERENCES

Moore, Henry L., *Laws of Wages*. ("Introduction" gives an excellent statement of the scientific approach to economic problems, followed by applications of the methods of correlation to economic phenomena.)

Elderton, W. P., *Frequency Curves and Correlation* (1906). *Addendum* (1917).

Mills, F. C., "The Measurement of Correlation and the Problem of Estimation," *Journal of the American Statistical Association*, September, 1924. (A general measure of correlation, index of correlation.)

Ross, Frank A., *School Attendance in the United States in 1920*, Appendix A, Census Monograph No. 5, Bureau of the Census, Washington, 1924. (The method of partial and multiple correlation applied to school attendance.)

Yule, G. U., *An Introduction to the Theory of Statistics*, 6th ed., chap. 12. (Correlation of more than two variables — multiple and partial correlation.)

Bowley, A. L., *Elements of Statistics*, 4th ed., Part II, chaps. 6, 7, and 8. (Partial and multiple correlation are treated in chapter 8.)

Pearl, Raymond, *Medical Biometry and Statistics*, chap. 15. (Multiple and partial correlation of more than two variables.)

Kelley, T. L., *Statistical Method*, chaps. 8, 10, and 11. (Multiple correlation treated by the same author in chapter 9 of the *Handbook of Mathematical Statistics*.)

Pearson, Karl, "On the General Theory of Skew Correlation and Non-Linear Regression," *Drapers' Company Research Memoirs:* Biometric Series II, London, 1905. (The correlation ratio.)

Blakeman, J., "On Tests for Linearity of Regression in Frequency Distributions," *Biometrika*, vol. IV, 1905, p. 332.

[Details as to publisher and date of publication of the readings and references given at the close of each chapter are found in the alphabetical list in Appendix A.]

~~Quantitative~~ scale
measurement

CHAPTER XIII

THE TIME SERIES — HISTORICAL STATISTICS

THE time factor has been repeatedly emphasized in the preceding pages as fundamental in the classification and interpretation of certain types of statistical data. In Chapter III social and economic phenomena were described as continually changing. Quantitative records taken at a specific point of time reveal relations in cross-section; records located at successive periods of time measure changes. Business activity is increasing or decreasing, but is rarely, if ever, static. By statistical methods norms may be established but these are usually trends in the upward or downward direction. The trend of the general mortality rate is downward, measuring the increasing control over health conditions. In Chapter IV the time series was contrasted with other types and such a classification was defended on the ground that appropriate methods must be devised for the treatment of different types of quantitative data. Index numbers have other important uses, but in Chapter X they have been described as statistical devices useful for measuring change over a period of time. For the most part, the preceding chapters have been devoted to general statistical methods and to those found to be appropriate for the treatment of variables in magnitude — frequency distributions. *The present chapter sets forth the elementary methods particularly adapted to the time series.*

time series [illegible] forecast [illegible] it will happen in future

CHARACTERISTICS PECULIAR TO THE TIME SERIES

The magnitudes in a time series, instead of being merely variables in amount as in the frequency distribution, apply to definite intervals of time — a week, a month, a quarter, a year. Each item is itself an aggregate, an average or a relative number, as, for example, the monthly or annual immigration, the average wholesale price of wheat, or the index number of food prices at retail. As a rule, the consecutive time intervals should be equal and *when the items are shown graphically the time series has a characteristic form.*

This form is the resultant of different types of movements due to natural, economic and social forces which at any moment may be operating in the same or in opposite directions and with varying intensities. Analysis of the time series into its principal elements is required.

(1) *The secular or long-time trend.* Over a considerable period of time

many social and economic phenomena exhibit a continuous tendency to grow or to decline. For example, population increases decade by decade while birth-rates decrease; with increasing business activity the volume of credit transactions moves steadily upward. The movement may be retarded over shorter periods of time or even reversed, but the general trend is upward or downward, due to the operation of persisting and relatively permanent forces.

(2) *Seasonal fluctuations.* In contrast to the long-time changes are those which occur within the limits of a single year. For many types of data there exists a characteristic variation from month to month. Certain trades are known as seasonal because of the wide variation in the amount of activity and the numbers employed; infant mortality rises to a peak in the months of July and August; the volume of local telephone traffic falls off in the summer months and increases again in the autumn; the demand for credit facilities and railway cars increases at the time of moving the crops to market. Weekly or monthly data and their analysis are necessary to reveal these seasonal movements, their characteristic forms and their causes.

The secular and the seasonal movements go on at the same time but not necessarily in the same direction. In business it is common practice to compare the facts of the current month with the corresponding data for the preceding month and for the same month of the preceding year. Such comparisons must be viewed with caution because, while they take account of seasonal changes, they frequently do not make allowance for the long time trend which proceeds normally from year to year. For example, if the volume of production of a basic raw material, pig iron, has been increasing at the average rate of five per cent annually over a period of years, then the current month's production should be about five per cent greater than that for the corresponding month last year in order to indicate equally prosperous conditions. The seasonal influences still operate, but on a higher level from year to year.

(3) *Cyclical movements.* These changes are best illustrated by the recurring, wave-like increases and decreases in business activity, which occur at more or less regular periods of time and which constitute the business cycle of prosperity and depression. Such short-time fluctuations may be accompanied also by seasonal variations, and may themselves be repeated again and again during the longer period of the secular trend of growth or decline. At any point in a given time series these various movements may coincide in direction or may take place in opposite directions. For example, the amount of bituminous coal mined in the United States in 1907, which represented a peak of production in a

prosperous period, was less than the amount mined in 1914, a period of depression. The level of the depression of 1914 was higher than the prosperity level of 1907, due to the growth of the industry during the intervening period.

We may wish to compare two time series. The long-time trends may move together, while the short-time fluctuations may be unrelated or moving in opposite directions. On the other hand, the secular movement of one series may be upward, as the volume of pig-iron production, 1903–1914, and the trend of a related series may be downward, as interest rates on 60–90 day commercial paper for the same period, while the short-time movements of both related series take place repeatedly in the same direction. Obviously, if we wish to measure the relationship between such series it will prove useless or misleading to compute a measure of correlation between pairs of the actual yearly or monthly items as they appear in the original data, because all the factors described influence these quantities. It would be impossible to interpret the results. *Therefore, for the purpose of comparison and study of the cyclical movements in related time series, undisturbed by other fluctuations, some method must be employed to eliminate the influence of both seasonal and secular movements in each series before the cycles are compared or correlated.*

(4) *Residual fluctuations.* Such movements might be caused by war, strikes, a great calamity such as fire or flood, or by governmental action. If these irregular fluctuations are to be eliminated from the original data a procedure adapted to the particular facts must be worked out for each situation as it arises. Therefore, this type of variation will be disregarded in our present discussion of methods.

(5) *The concept of lag and lead in a time series.* Suppose the problem is to investigate the relation of the number of deaths under one year of age to the changes in the mean monthly temperature. It is a familiar fact that more babies die during the hot summer months. We may discover that the mean temperature rises to a maximum in July and then falls off, whereas the number of deaths does not reach the highest point until some time in August. Apparently the increase in deaths occurs after the rise in temperature, or, in other words, *lags behind the temperature changes.* If we superimpose the two curves on the same sheet a closer correspondence in the movements will be observed by plotting each month's deaths one month earlier than they actually occurred, August deaths opposite July temperatures. *This assumes a lag of one month in the death series or a lead of one month in the temperature series.*

It is a familiar observation that as commodity prices rise in a period of prosperity wages do not respond at once, but only after a lapse of time.

We designate this period of time between changes in one time series and changes in a related series as the lag of the one series or the lead of the other. For example, wage changes may lag behind price changes, or the volume of pig-iron production may anticipate a change in interest rates.

Comparison of time series involves pairing items definitely related in time but it is not necessary to pair only those items which refer to the same month or year. Inspection of a diagram may suggest the most probable period of lag which should be tested by moving one of the series forward or backward until the closest correspondence in the variations is secured. *In business the discovery of leads and their measurement furnishes a basis for forecasting changes months before they actually occur.*

THE DETERMINATION OF SECULAR TREND

There are several methods of determining the trend or long-time movement as distinguished from the cyclical or short-time changes. Figures 36 and 37 present, for the period 1903–14, the variations in

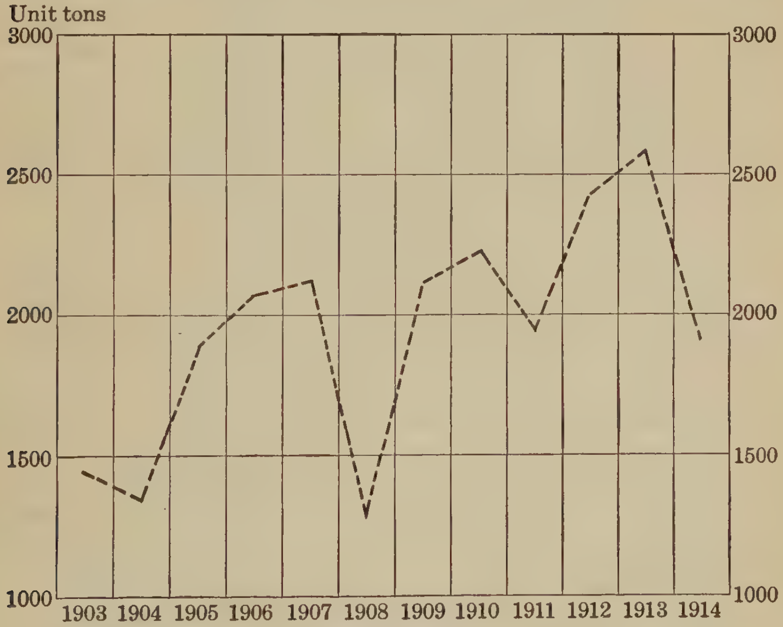


FIG. 36. AVERAGE MONTHLY PRODUCTION OF PIG IRON IN THE UNITED STATES, 1903–1914[1]

(Unit = 1000 long tons. Data from Table 60.)

[1] In this diagram and many of those which follow in this chapter the zero base line is omitted to save space in plotting and because the zero line is not essential in the interpretation. In this case the bottom line of the diagram is made the same weight as the other lines of the background.

average monthly production of pig iron in the United States and in the average annual interest rates on 60–90 day commercial paper in New York City. The data are found in Table 60, columns (1) and (6).

THE MOVING AVERAGE

Pig iron constitutes the raw material for a great variety of important manufactured products, and the interest rate on commercial paper is a sensitive index of business activity. From Figures 36 and 37 it is clear that the short-time movements are more or less closely associated and in the same direction. What is true of the general movement or trend of each series for the entire period? Does the long-time movement exhibit a growth, a decline, or does it move forward on about the same level? Inspection shows that the two series differ in their long-time movements.

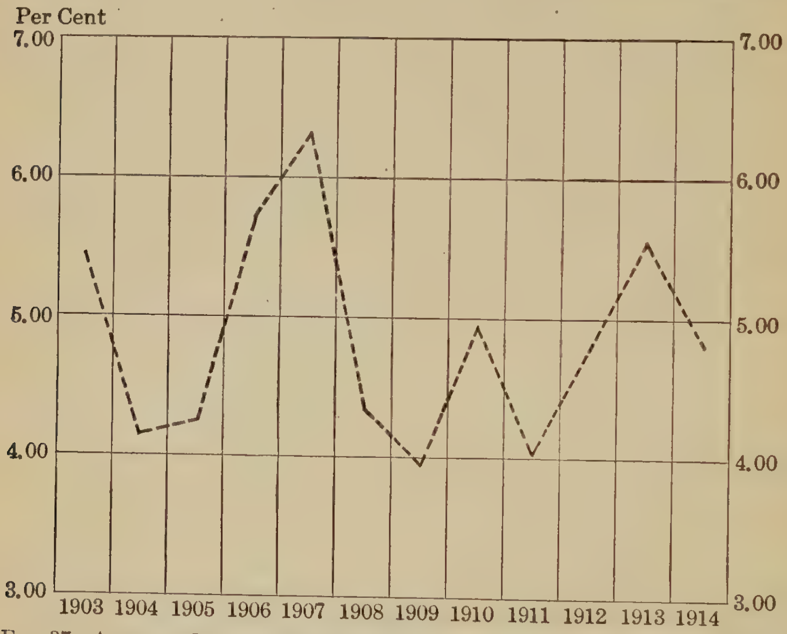


Fig. 37. Average Rate of Interest on 60 to 90 Day Commercial Paper, New York City, 1903–1914

(Unit = one per cent. Data from Table 60. See footnote to Fig. 36.)

In order to reveal the trend in the series in a simple manner it is possible to smooth out the short-time fluctuations. To determine the general movement the items may be averaged for such a period of time (3, 4, 5, 7 years), as will include both high and low values in the short-time fluctuations — *at least a completed cycle.* In order to obtain a mean value for each year or other unit of time in succession the method of a *moving average* has been utilized. Table 60 presents several moving averages for pig-iron production and interest rates, each employing a different time interval for combining the original items. *The moving average is not necessarily the best method of determining the secular trend, and other methods will be discussed later.*

TABLE 60. MOVING AVERAGES OF PIG-IRON PRODUCTION AND INTEREST RATE

Year	Pig-Iron Production (Unit = 1000 long tons)					Interest Rates on 60–90 Day Commercial Paper, New York City (Unit = 1 per cent)				
	Average monthly production (1)	Three-year moving average (2)	Four-year moving average (3)	Four-year moving average centered[a] (4)	Five-year moving average (5)	Average annual interest rates[b] (6)	Three-year moving average (7)	Four-year moving average (8)	Four-year moving average centered[a] (9)	Five-year moving average (10)
1901	1282					4.28				
1902	1435					4.92				
			1378					4.72		
1903	1452	1410		1453	1479	5.47	4.87		4.74	4.66
			1528					4.75		
1904	1344	1559		1607	1636	4.21	4.69		4.84	4.94
			1686					4.94		
1905	1882	1764		1768	1771	4.40	4.76		5.05	5.22
			1850					5.16		
1906	2066	2019		1845	1741	5.68	5.48		5.19	5.01
			1840					5.21		
1907	2109	1826		1869	1895	6.36	5.47		5.15	4.96
			1898					5.10		
1908	1302	1842		1920	1966	4.38	4.91		5.02	5.08
			1941					4.93		
1909	2116	1885		1920	1942	3.98	4.45		4.64	4.75
			1900					4.35		
1910	2237	2099		2043	2009	5.00	4.34		4.39	4.43
			2186					4.44		
1911	1944	2210		2242	2261	4.03	4.59		4.64	4.67
			2297					4.84		
1912	2448	2317		2257	2222	4.74	4.79		4.82	4.83
			2218					4.79		
1913	2560	2310		2284	2269	5.60	5.04		4.71	4.52
			2350					4.64		
1914	1921	2318		2451	2531	4.78	4.61		4.48	4.40
			2551					4.32		
1915	2472					3.45				
1916	3252					3.43				

[a] The centering of values is explained on page 312.

[b] *The Review of Economic Statistics*, Committee on Economic Research, Harvard University, Preliminary Volume I, pp. 98 and 99.

In column (2) of Table 60 the quantities are three-year averages of the items in column (1), obtained as follows: $\frac{1435 + 1452 + 1344}{3} = 1410$; $\frac{1452 + 1344 + 1882}{3} = 1559$; $\frac{1344 + 1882 + 2066}{3} = 1764$, etc. One year is omitted each time and another year is added. The result is a series of *moving averages* in column (2), each of which represents a period of three years, one year on each side of that opposite which it is entered. In Figure 36 the original data of column (1) are plotted in the middle of successive years. Likewise, the moving averages of column (2) are plotted in Figure 38, each in the middle of a three-year period, which locates the quantities in the middle of successive years, opposite the original data.

The quantities in column (3) are obtained by averaging four items of column (1) instead of three, $\frac{1282 + 1435 + 1452 + 1344}{4} = 1378$, etc. The resulting quantity, 1378, should not be entered opposite the year 1903, but between 1902 and 1903, having two years on each side, and in like manner all values in column (3) are located. But this procedure means that the *four-year moving averages* in column (3) do not represent the same point of time as the quantities in columns (1) and (2), by a difference of a half-year. This proves inconvenient if we wish to plot the values of column (3) along with those of column (1) since the plotted points are not located opposite each other for comparison. Besides, we cannot subtract the values of column (1) from those of column (3) to ascertain the difference between the original data and a four-year moving average representing the trend.

This difficulty can be removed in a simple manner. In Column (4) the quantities are obtained by taking an additional two-year moving average of the items in column (3), $\frac{1378 + 1528}{2} = 1453$, etc. Now the quantity 1453 is entered opposite the year 1903 and would be plotted opposite the middle of the year, because the quantities averaged to obtain it, 1378 and 1528, are located at the beginning and the end of 1903. In a similar manner all the items in column (4) are obtained.[1] This procedure is called *centering the values at the middle of the year.*

[1] In averaging two numbers the result is either an integer as in the illustration, 1453; or the result is an integer and the fraction ½, for example, from column (3), $\frac{1941 + 1900}{2} = 1920\frac{1}{2}$, or $\frac{2186 + 2297}{2} = 2241\frac{1}{2}$, etc. The usual procedure in eliminating fractions is to count ½ at the next higher integer. In the present case this would *cumulate*

This illustration emphasizes the advantage of taking, if possible, an odd number of years for the interval of the moving average. If this is done the resulting values will be located in the middle of the successive years, as shown in columns (2) and (5), opposite the original data of column (1). Any even number of years will result as in columns (3) and (8).

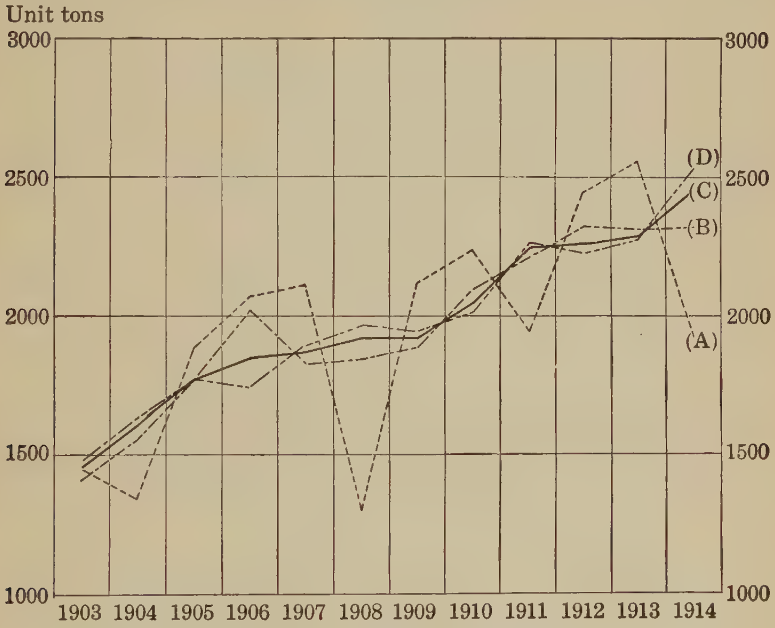


FIG. 38. MOVING AVERAGES OF MONTHLY PIG-IRON PRODUCTION IN THE UNITED STATES, 1903–1914

(Unit = 1000 long tons. Data from Table 60. See footnote Fig. 36.)

Key: (A) = Original data.
(B) = Three-year moving average.
(C) = Four-year moving average, centered.
(D) = Five-year moving average.

The same difficulty arises in taking a twelve-month moving average to smooth out seasonal fluctuations. This will be discussed in a later section.

Figure 38 presents the results of experimentation with moving averages of different intervals (3, 4 and 5 years), plotted from data in columns (1), (2), (4) and (5) of Table 60. The three-year moving average

the error because the fraction, when it appears at all, is always ½. In eliminating fractions the principle is to make the amounts dropped and added balance each other. In order to follow this principle in a two-year moving average, the fraction ½ must be counted to the next higher integer and dropped entirely at alternate occurrences. This has been done in columns (4) and (9) of Table 60.

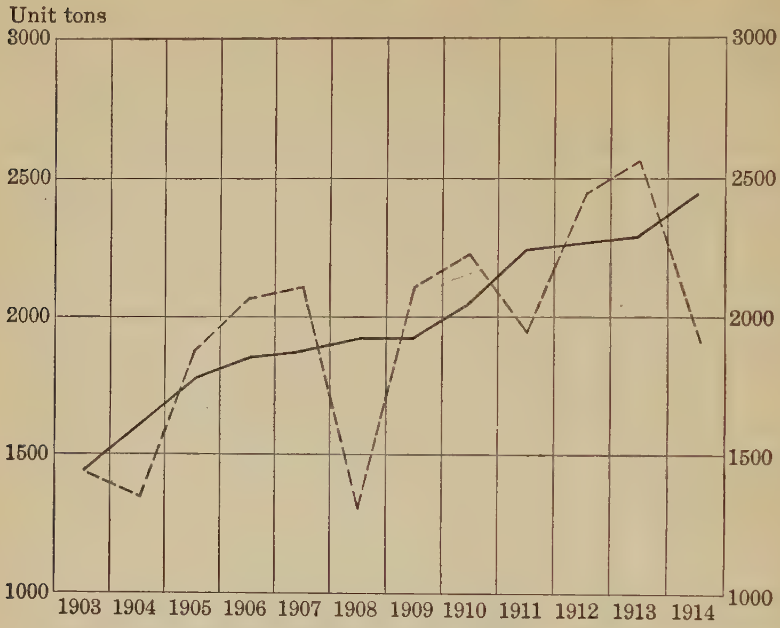


FIG. 39. AVERAGE MONTHLY PRODUCTION OF PIG IRON, 1903–1914 (DOTTED LINE), AND A FOUR-YEAR MOVING AVERAGE, REPRESENTING THE TREND (CONTINUOUS LINE)

(Unit = 1000 long tons. Data from Table 60, columns (1) and (4). See footnote Fig. 36.)

fails to smooth out the short-time fluctuations of the early part of the period. The five-year average establishes a more satisfactory trend line but presents certain logical contradictions. It is not expected that the moving average will follow a straight line, but it will be observed that in this case *the five-year line curves in the opposite direction from the original data at each of the short-time movements.* The four-year average, centered at the middle of the year, improves this situation, but it is not so simple in its computation. In Figures 39 and 40 the four-year moving average is used to describe the trend of the data because it seems to smooth out the short-time fluctuations in a manner somewhat better than the other moving averages.[1]

The moving average as a zero line from which to measure short-time fluctuations. With the *lines of secular trend* established for the period 1903–1914 by a four-year moving average which smooths out the short-

[1] A seven-year moving average was tried also but rejected because the results were not appreciably better, and this average requires additional data at the beginning and end of the period. Besides, the inclusion of the war years at the end of the period tends to elevate the slope of the trend.

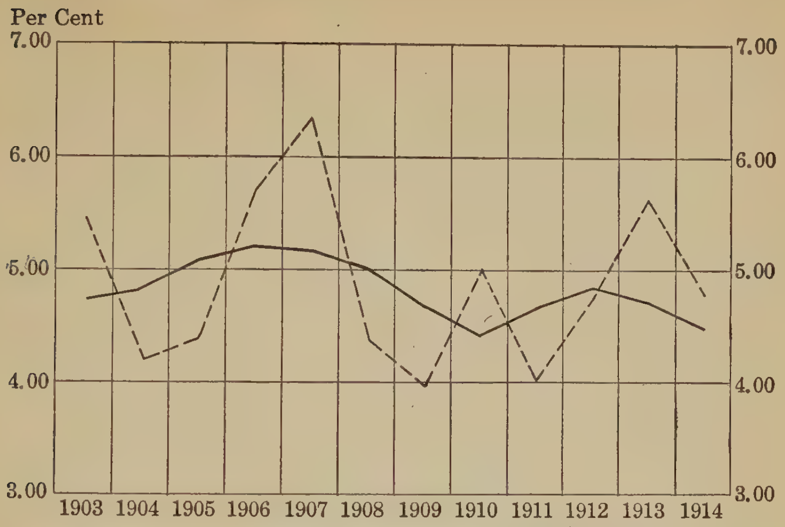


FIG. 40. AVERAGE RATE OF INTEREST ON 60 TO 90 DAY COMMERCIAL PAPER, NEW YORK CITY (DOTTED LINE), AND A FOUR-YEAR MOVING AVERAGE REPRESENTING THE TREND (CONTINUOUS LINE)

(Unit = one per cent. Data from Table 60, columns (6) and (9). See footnote Fig. 36.)

time fluctuations, the *cyclical variations* from these trends are shown in Table 61.

Column (3) in Table 61 shows how much the production deviates

TABLE 61. CYCLICAL VARIATIONS MEASURED FROM MOVING AVERAGES

YEAR	PIG-IRON PRODUCTION			INTEREST RATES		
	Average monthly production (unit = 1000 long tons) (1)	Four-year moving average = 0 (2)	Deviations of actual from average (3)	Average annual interest rate (unit = 1 per cent) (4)	Four-year moving average = 0 (5)	Deviations of actual from average (6)
1903	1452	1453	− 1	5.47	4.74	+ .73
1904	1344	1607	−263	4.21	4.84	− .63
1905	1882	1768	+114	4.40	5.05	− .65
1906	2066	1845	+221	5.68	5.19	+ .49
1907	2109	1869	+240	6.36	5.15	+1.21
1908	1302	1920	−618	4.38	5.02	− .64
1909	2116	1920	+196	3.98	4.64	− .66
1910	2237	2043	+194	5.00	4.39	+ .61
1911	1944	2242	−298	4.03	4.64	− .61
1912	2448	2257	+191	4.74	4.82	− .08
1913	2560	2284	+276	5.60	4.71	+ .89
1914	1921	2451	−530	4.78	4.48	+ .30

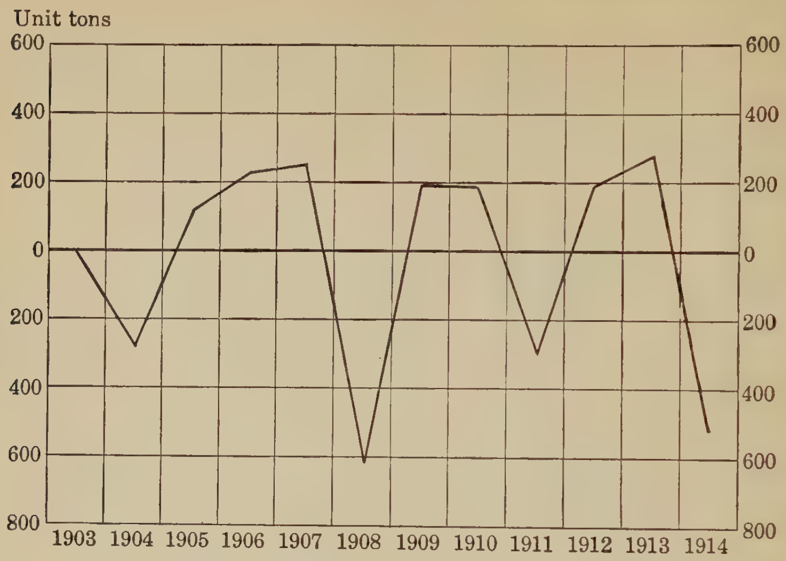


FIG. 41. ANNUAL FLUCTUATIONS IN PIG-IRON PRODUCTION ABOUT THE TREND, 1903–1914

Cycles measured from four-year moving averages, representing the trend as zero. (Unit = 1000 long tons. Data from Table 61, column (3). Deviations from zero.)

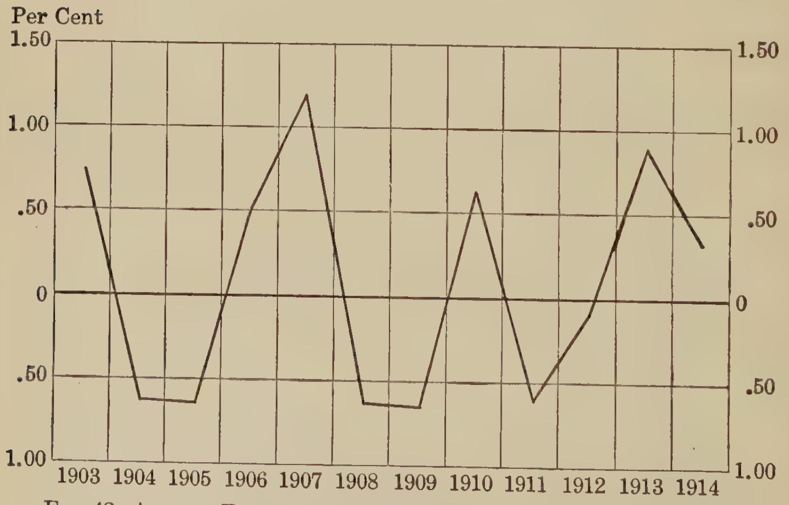


FIG. 42. ANNUAL FLUCTUATIONS IN INTEREST RATE ON 60 TO 90 DAY COMMERCIAL PAPER, NEW YORK CITY, ABOUT THE TREND, 1903–1914

Cycles measured from four-year moving averages, representing the trend as zero. (Unit = one per cent. Data from Table 61, column (6). Deviations from zero.)

above or below the trend at any given year. The trend, column (2), describes a *normal line from which to measure deviations.* When the sign is minus, production is below normal, and the opposite is indicated by the plus sign. Column (6) shows the same for interest rates.

In Figures 41 and 42 the *four-year moving average is shown as a straight line,* designated as zero on the vertical scale, because it is used as a norm at each year from which the plus and minus deviations are measured, and because it is desired to show only the short-time fluctuations about the secular trend. *In other words, this procedure eliminates the secular trend while we examine only the cyclical movements in the series.*

We have measured these fluctuations in the original units, tons and per cents of interest rate. Frequently, they are measured as *percentage variations from the trend line,* always using the moving averages or trend-line values as the base for the calculation of the percentages. This procedure reduces all variations to a common unit, percentages of the ordinates of trend, regardless of the variety of units in the original data. The vertical scales can be expressed in percentages for the various series whose fluctuations about the trend lines are compared. The curves are similar to those shown in Figures 41 and 42. This method is illustrated in a later section of this chapter.

Correlation of the short-time fluctuations about the trends of two series. Inspection of Figures 39 and 40 shows a decided upward movement in the secular trend of pig-iron production over the period 1903–1914, while the trend of interest rates is slightly in the opposite direction. On the other hand, the short-time fluctuations in the two series appear to move in the same direction with a fair degree of regularity. If we desire to describe and to measure the relationship between the short-time fluctuations alone in the two series, it is logically clear that the secular trends should be eliminated first. Otherwise, the relationships of the long and short-time changes will be confused and may neutralize each other. As evidence on this point we shall correlate the series in their original form (Table 62), and then correlate only the short-time fluctuations, after the secular trends have been eliminated by the method of a moving average, as shown in Figures 41 and 42.

The correlation is apparently low. Will the coefficient be changed if the secular trends are eliminated by the device of a four-year moving average before correlating the fluctuations? The data for Table 63 are taken from Table 61, columns (3) and (6), where the deviations are given from the moving average as zero.

The coefficient proves to be considerably higher when the secular trends which move in opposite directions have been eliminated and

TABLE 62. CORRELATION OF FLUCTUATIONS IN PIG-IRON PRODUCTION AND INTEREST RATES (1903–1914)

(Original series, without eliminating secular trends)

YEAR	PIG-IRON PRODUCTION (unit = 1000 tons)	DEVIATION FROM AVERAGE FOR 12 YEARS		INTEREST RATES (unit = 1 per cent)	DEVIATION FROM AVERAGE FOR 12 YEARS		*xy* PRODUCTS (7)	
	X	x	x^2	Y	y	y^2		
	(1)	(2)	(3)	(4)	(5)	(6)	+	−
1903	1452	−496	246,016	5.47	+ .58	.3364		287.68
1904	1344	−604	364,816	4.21	− .68	.4624	410.72	
1905	1882	− 66	4,356	4.40	− .49	.2401	32.34	
1906	2066	+118	13,924	5.68	+ .79	.6241	93.22	
1907	2109	+161	25,921	6.36	+1.47	2.1609	236.67	
1908	1302	−646	417,316	4.38	− .51	.2601	329.46	
1909	2116	+168	28,224	3.98	− .91	.8281		152.88
1910	2237	+289	83,521	5.00	+ .11	.0121	31.79	
1911	1944	− 4	16	4.03	− .86	.7396	3.44	
1912	2448	+500	250,000	4.74	− .15	.0225		75.00
1913	2560	+612	374,544	5.60	+ .71	.5041	434.52	
1914	1921	− 27	729	4.78	− .11	.0121	2.97	
Average for 12 years	1948	......	....	4.89				
Summations .		....	1,809,383	....	....	6.2025	+1575.13	−515.56

$$\sigma_X = \sqrt{\frac{1,809,383}{12}} = 388 \qquad \sigma_Y = \sqrt{\frac{6.2025}{12}} = .72$$

$$r = \frac{\Sigma xy}{N\sigma_X\sigma_Y} = \frac{+1575.13 - 515.56}{12 \text{ times } 388 \text{ times } .72} = \frac{+1059.57}{3352.32} = +.32$$

only the cyclical fluctuations are related (+.32 compared with +.47). Evidently the long-time changes, until eliminated by the moving average or some other statistical device, interfere with the measurement of the relationship between the short-time movements of the two series.

Limitations of the moving average. The moving average and the moving total [1] have long been used to describe trends in historical data. The moving average is sensitive to a change in the direction of the trend over long periods, for example, in describing the movement of wholesale prices in the United States since the Civil War. In such a situation it takes the place of the more complicated methods of curve fitting.

[1] The moving total is similar to the moving average except that the sum of the items is not divided by their number. For a description of its uses and graphic representation, see Karl G. Karsten: *Charts and Graphs*, pp. 228–34.

TABLE 63. CORRELATION OF FLUCTUATIONS IN PIG-IRON PRODUCTION AND INTEREST RATES

(Secular trends eliminated)

YEAR	PIG-IRON PRODUCTION DEVIATIONS FROM MOVING AVERAGE (unit = 1000 tons)		INTEREST RATES DEVIATIONS FROM MOVING AVERAGE (unit = 1 per cent)		PRODUCTS xy (5)	
	x (1)	x^2 (2)	y (3)	y^2 (4)	+	−
1903	− 1	1	+ .73	.5329		.73
1904	−263	69,169	− .63	.3969	165.69	
1905	+114	12,996	− .65	.4225		74.10
1906	+221	48,841	+ .49	.2401	108.29	
1907	+240	57,600	+1.21	1.4641	290.40	
1908	−618	381,924	− .64	.4096	395.52	
1909	+196	38,416	− .66	.4356		128.70
1910	+194	37,636	+ .61	.3721	118.34	
1911	−298	88,804	− .61	.3721	181.78	
1912	+191	36,481	− .08	.0064		15.20
1913	+276	76,176	+ .89	.7921	245.64	
1914	−530	280,900	+ .30	.0900		159.00
	[a] +1432 −1710	1,128,944	[a] +4.23 −3.27	5.5344	+1505.66	−377.73

$$c_X = \frac{-1710 + 1432}{12} = -23 \qquad c_Y = \frac{+4.23 - 3.27}{12} = +.08$$

$$\sigma_X = \sqrt{\frac{1,128,944}{12}} - (-23)^2 = 306 \qquad \sigma_Y = \sqrt{\frac{5.5344}{12}} - (+.08)^2 = .67$$

$$r = \frac{\dfrac{+1505.66 - 377.73}{12} - (-23 \text{ times } +.08)}{306 \text{ times } .67} = \frac{93.99 + 1.84}{205.02} = +.47$$

[a] Unlike the situation in Table 62, when the moving average is used the sum of the deviations plus and minus does not equal zero. Therefore the differences must be noted and a correction factor must be computed, as in the case of a guessed average in a frequency distribution (see Columns (1) and (3) of Table 63).

However, the following disadvantages in the use of a moving average should be noted:

(1) It is easily affected by extreme variations, as would be illustrated by extending the moving average of pig-iron production or interest rates beyond 1914 into the war period.

(2) It is not an easy task to determine the best period for such an average in months or years (3, 5, 7, 9 years). Comparison of the curves in Figure 38 illustrates this difficulty.

(3) It does not cover the beginning or the end of the period under examination without estimates. Frequently, the latest data are most important, especially when it is desired to project the trend into the future.

(4) When used for purposes of correlating two or more series it requires *a correction factor*, as shown in Table 63, because the sum of the deviations above and below the moving average does not equal zero. *In this respect it is similar to a guessed average in a frequency distribution.* Furthermore, except in cases where there is good reason to think the trend is linear, the moving average should be employed with caution since it is likely to introduce an element of spurious correlation. If the trend is really in the form of a parabola the moving average not only smooths the data but tends to change the form of the trend.

FITTING A STRAIGHT LINE TO HISTORICAL DATA DESCRIBING THE SECULAR TREND

The original data on pig-iron production have been plotted in Figure 36, and inspection suggests that a *straight line* will describe the secular trend or growth during the period 1903–1914 as well or better than the four-year moving average shown in Figure 39, or any more complicated form of curve. The straight line hypothesis means that the growth is described by a *constant average annual increase* in tons year-by-year over the period of 12 years. Values are located on the straight line at each year, from which the cyclical variations may be measured in the same manner as from the moving average already described. *The object is the same — to eliminate the secular trend and to examine the short-time fluctuations.* The straight line may be fitted to a historical series by several methods.

(1) *Method of inspection.* The line may be located by the eye, by the use of a ruler, or with the aid of two thumb tacks and a thread. In any case, it is so located as to approximate in the best manner the trend of all the points plotted from the original data. Sometimes the average of the entire series is plotted at the middle of the period and the line is passed through this point, fitting as closely as possible the other points plotted at each year. For example, the average monthly production of pig iron for 12 years, 1903–1914, is 1948 unit-tons. This value may be plotted as an ordinate located between 1908 and 1909, with a six-year period on either side. Through this point the straight line is drawn, as illustrated in Figure 43.

(2) *Method of semi-averages.* The pig-iron production series, for example, 1903–1914, is divided into two equal parts of 6 years each. The average monthly production 1903–1908 is 1693 unit-tons, plotted as an ordinate between 1905 and 1906, the middle of the six-year period. Likewise, the average for the period 1909–1914 is 2204 unit-tons, plotted between 1911 and 1912. The straight line is passed through these two points which describe each half of the series, as shown in Figure 43. It is

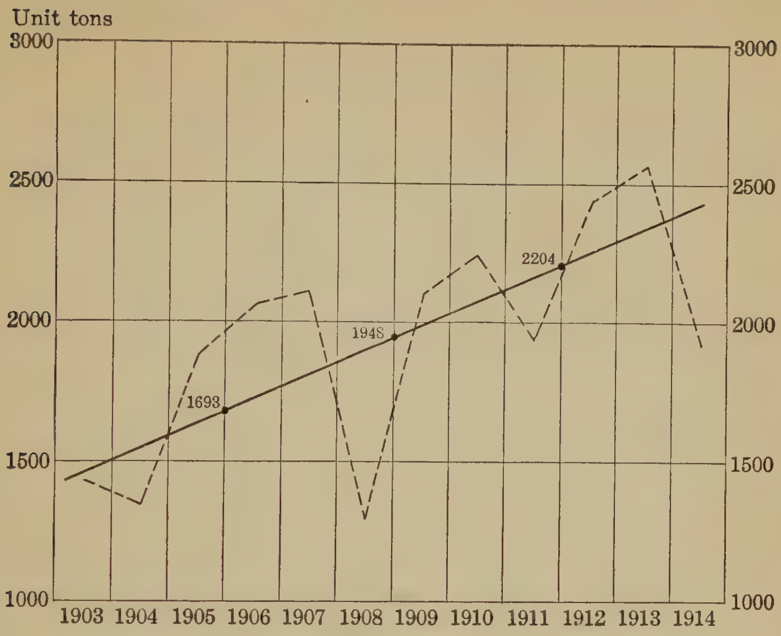


FIG. 43. A STRAIGHT LINE TREND FITTED BY THE METHOD OF SEMI-AVERAGES

Average monthly production of pig iron, 1903–1914, in two periods of six years each. (Unit = 1000 long tons. See footnote to Fig. 36.)

clear that this line located by the method of semi-averages and describing the growth of pig-iron production also passes through the point representing the average for the entire series, 1948 unit-tons.

(3) *Method of correlating with time. A regression line of Y on X may be used to describe the straight-line trend.* The method of establishing this regression line is already familiar from the examples presented in the last chapter. In the present case the X variable is time (1903–1914), and the Y variable is pig-iron production. We shall correlate the movements of these two variables. Since the method is a little clearer when an *odd number of years* is taken, the first correlation will be for the period 1903–1915, Table 64.

Regression of Y on X — the secular trend.

$$y = r\left(\frac{\sigma_Y}{\sigma_X}\right)x, \text{ or } y = +.70\left(\frac{398}{3.74}\right)x, \text{ or } y = 74\,x.$$

This equation describes the trend of pig-iron production 1903–1915, on the hypothesis that it is a straight line passing through the average produc-

TABLE 64. CORRELATION OF THE CHANGES IN PIG-IRON PRODUCTION WITH TIME, 1903–1915

(13 years — an odd number)

TIME (X)			PIG-IRON PRODUCTION (Y)				
YEAR	Deviations from mid-year 1909		Pig-iron production	Deviations from average 1989		xy Products (7)	
X (1)	x (2)	x^2 (3)	Y (4)	y (5)	y^2 (6)	+	−
1903	−6	36	1452	−537	288,369	3,222	
1904	−5	25	1344	−645	416,025	3,225	
1905	−4	16	1882	−107	11,449	428	
1906	−3	9	2066	+ 77	5,929		231
1907	−2	4	2109	+120	14,400		240
1908	−1	1	1302	−687	471,969	687	
1909	0	0	2116	+127	16,129	0	
1910	1	1	2237	+248	61,504	248	
1911	2	4	1944	− 45	2,025		90
1912	3	9	2448	+459	210,681	1,377	
1913	4	16	2560	+571	326,041	2,284	
1914	5	25	1921	− 68	4,624		340
1915	6	36	2472	+483	233,289	2,898	
Mid-year = 1909		182	Average = 1989 (1988.7)		2,062,434	+14,369	−901

$$\sigma_X = \sqrt{\frac{182}{13}} = 3.74 \text{ years; } \sigma_Y = \sqrt{\frac{2{,}062{,}434}{13}} = 398.$$

$$r = \frac{+14{,}369 - 901}{13 \times 3.74 \times 398} = \frac{+13{,}468}{19{,}351} = +.70$$

tion for the entire period 1903–1915, which is 1989 units, plotted in the middle of the year 1909 as origin. The equation means that each year of time change (x) is accompanied *on the average* by a change of 74 units, of 1000 tons each, in the amount of pig-iron production (y).

We shall now show the method of correlating with time when the period is an *even number of years*, 1903–1914, which has been the period used in the preceding sections of this chapter (Table 65).

It will be noted that column (2) in Table 65 requires some explanation. Whenever an even number of time units, in this case 12 years, is used, the mid-point of the entire series falls between two years. On this account, in column (2) the *unit of deviation is made a half-year*, measured from the point between the years 1908 and 1909 as origin, in order that the units of deviation may fall opposite the middle of the years up and down the column. The first minus step-deviation, opposite the middle of 1908, is $\frac{1}{2}$ year; the second, opposite 1907, is $\frac{3}{2}$ years, etc. Using

Table 65. Correlation of the Changes in Pig-Iron Production with Time, 1903–1914

(12 years — an even number)

Year	Time (X)		Pig-Iron Production (Y)					Slope = 72
	Deviations from mid-point 1908–09 (unit = half-year)		Pig-Iron Production	Deviations from average 1948 units		*xy* Products (7)		Trend line values. Origin 1908–1909
X (1)	x (2)	x^2 (3)	Y (4)	y (5)	y^2 (6)	+	−	(8)
1903	−11	121	1452	−496	246,016	5,456		1552
1904	− 9	81	1344	−604	364,816	5,436		1624
1905	− 7	49	1882	− 66	4,356	462		1696
1906	− 5	25	2066	+118	13,924		590	1768
1907	− 3	9	2109	+161	25,921		483	1840
1908	− 1	1	1302	−646	417,316	646		1912
Origin			**1948**					**1948**
1909	1	1	2116	+168	28,224	168		1984
1910	3	9	2237	+289	83,521	867		2056
1911	5	25	1944	− 4	16		20	2128
1912	7	49	2448	+500	250,000	3,500		2200
1913	9	81	2560	+612	374,544	5,508		2272
1914	11	121	1921	− 27	729		297	2344
		572	Average = 1948 (1948.4)		1,809,383	+22,043	−1390	

$$\sigma_X = \sqrt{\frac{572 \div 4}{12}} = 3.45 \text{ years;}\quad \sigma_Y = \sqrt{\frac{1,809,383}{12}} = 388.$$

$$r = \frac{\dfrac{+22,043 - 1390}{2}}{12 \times 3.45 \times 388} = +.64$$

merely the numerators as step-deviations, minus and plus, we have the intervals as given in column (2) of Table 65. As a result of this procedure, in computing σ_X, the summation ($\Sigma x^2 = 572$) is in terms of half-year units each of which has been squared, column (3). To convert the summation into year units again, 572 is divided by 4 before being divided by the number of years in the entire series:

$$\sigma_X = \sqrt{\frac{572 \div 4}{12}} = 3.45 \text{ years.}$$

For the same reason the summation of the *xy* products is divided by 2 in the numerator of the *r* formula:

$$r = \frac{\dfrac{+22,043 - 1390}{2}}{12 \text{ times } 3.45 \text{ times } 388} = +.64$$

The secular trend 1903–1914. The secular trend of the twelve-year period may be described by the slope of the line of regression of Y on X, or the average amount of change in pig-iron production (y), corresponding to a unit change in time (x). The equation of this line is:

$$y = +.64 \left(\frac{388}{3.45}\right) x, \text{ or } y = 72\,x$$

The average annual growth in pig-iron production may be stated as 72 units of 1000 tons each, during the period 1903–1914.

Annual values on the trend line. In Table 65 the origin is at the mid-point of the series, between 1908 and 1909, where the average monthly production for the 12 years, 1948 units, is located. The other values, representing successive years, which are located on the straight line passing through the point representing 1948, may be calculated easily by substituting values of x, as shown in column (2) of Table 65, in the equation $y = 72\ x$, and by adding or subtracting these results to or from the central value 1948. For example,

$$y = \tfrac{1}{2} \text{ of } 72 = \pm\ 36 \text{ units} \qquad y = \tfrac{7}{2} \text{ of } 72 = \pm 252$$
$$y = \tfrac{3}{2} \text{ of } 72 = \pm 108 \text{ units} \qquad y = \tfrac{9}{2} \text{ of } 72 = \pm 324$$
$$y = \tfrac{5}{2} \text{ of } 72 = \pm 180 \text{ units} \qquad y = \tfrac{11}{2} \text{ of } 72 = \pm 396$$

It only remains to subtract these values of y in succession from 1948 to obtain the values for successive years below the mid-point of the series; and to add the same values successively to 1948 to obtain the values for the years above the mid-point. These values for each year on the trend line are recorded in column (8) of Table 65. Since the constant increment is 72 units each year, the line must be straight (Figure 44, page 326).

An abbreviation in the method of fitting the trend line. *The trend line is so located that the sum of the squares of the deviations of the actual data from it is a minimum, the deviations having been measured along the ordinates.*[1] (See Figure 44, distances from actual data to straight line, y_1, y_2, y_3, y_4, etc.) The student is already familiar with the function of r in determining the slope of this line by the use of the equation

$$y = r\left(\frac{\sigma_Y}{\sigma_X}\right) x$$

Let us see how this equation may be modified so as to determine the slope of the trend line directly from the data *without first computing* r.

[1] The principle of least squares has been utilized already in locating the regression line of Y on X by the method of correlating pig-iron production with time.

In the simple equation describing the straight line the slope is the value of m. In terms of the regression equation of Y on X,

$$m \text{ (slope)} = r\left(\frac{\sigma_Y}{\sigma_X}\right).$$

Substituting $\frac{\Sigma xy}{N\sigma_X\sigma_Y}$ for r in the equation we have,

$$m \text{ (slope)} = \frac{\Sigma xy}{N\sigma_X\sigma_Y} \text{ times } \frac{\sigma_Y}{\sigma_X} = \frac{\Sigma xy}{N\sigma_X{}^2} \text{ (cancelling } \sigma_Y).$$

But, $\sigma_X{}^2 = \frac{\Sigma x^2}{N}$, and $N\sigma_X{}^2 = \left(\frac{\Sigma x^2}{N} \text{ times } \frac{N}{1}\right) = \Sigma x^2.$

Therefore, $\frac{\Sigma xy}{N\sigma_X{}^2} = \frac{\Sigma xy}{\Sigma x^2} =$ *Slope of the trend line.*

The entire procedure of fitting the straight line is much simplified by the use of the equation, $m = \frac{\Sigma xy}{\Sigma x^2}$, and exactly the same results are obtained as those shown in Table 65 for the pig-iron production. Column (7) of Table 65 gives $\Sigma xy = +22043 - 1390 = +20653$, and column (3) gives $\Sigma x^2 = 572$. Therefore, the slope or average annual growth of pig-iron production $= \frac{20653 \div 2}{572 \div 4} = 72$, which is the same as that obtained by correlating with time and by locating the regression line of Y on X. The reason for dividing the numerator by 2 and the denominator by 4 is the same as that explained in detail following Table 65. To avoid fractions the units of deviation (x) are stated in half-years, column (2), and, therefore, *must be reduced to the annual basis.* Since the squares were used in column (3), it is necessary to divide the denominator (Σx^2) by 4.

The procedure for locating the values for each year on the line of secular trend is exactly the same as that explained after Table 65. If the sign is plus, as in this case, the slope is upward and to the right; if the sign is minus, the trend is downward and toward the right. On the hypothesis that a straight line fits the data, which must be determined for each time series and period described, this abbreviated method of correlating with time should be used in locating the line of secular trend.

We shall now apply this method to determine the trend of interest rates, 1903–1914, where the movement is downward, opposite in direction to that of pig-iron production, as shown by the moving average (Figure 40, page 315).

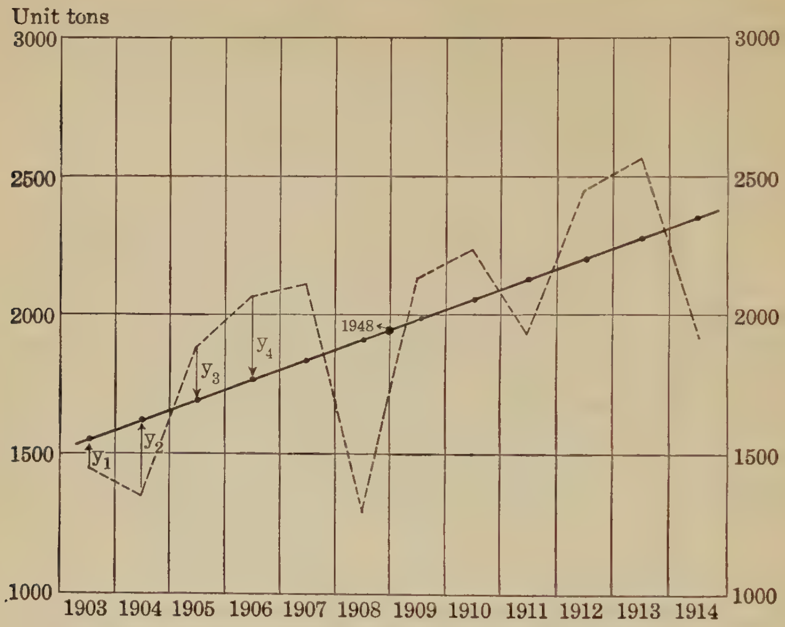


FIG. 44. SECULAR TREND OF PIG-IRON PRODUCTION, 1903–1914

Straight line fitted by the method of least squares or correlation with time. (Unit = 1000 long tons. Data from Table 65, column (8). See footnote Fig. 36. Equation to Trend Line is $y = +72x$, origin between 1908 and 1909.)

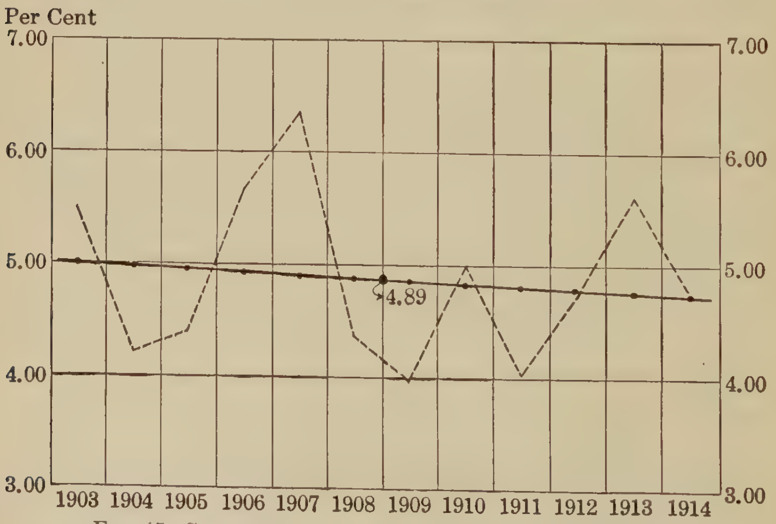


FIG. 45. SECULAR TREND OF INTEREST RATES, 1903–1914

Straight line fitted as in Fig. 44. (Unit = one per cent. Data from Table 66, column (7). Equation to Trend Line is $y = -.02x$, origin between 1908 and 1909.)

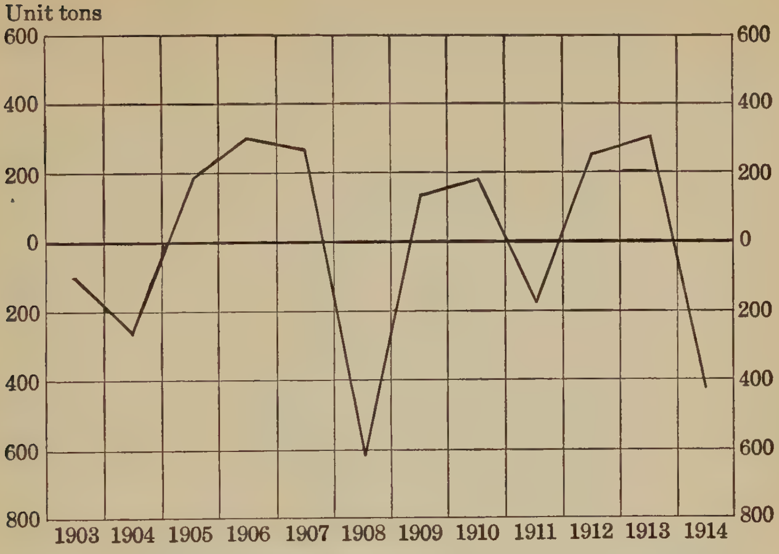


FIG. 46. CYCLICAL FLUCTUATIONS OF PIG-IRON PRODUCTION, 1903–1914

Deviations are plotted from the secular trend as zero. (Unit = 1000 long tons. Data from Table 67, column (3), page 329, in which the influence of secular trend is eliminated.)

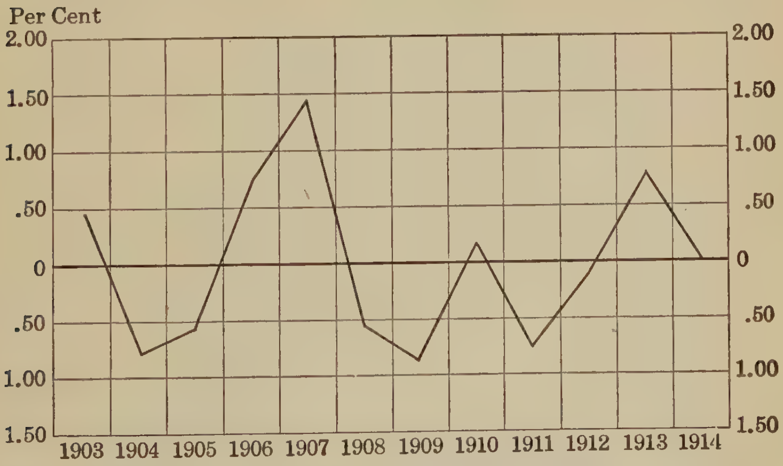


FIG. 47. CYCLICAL FLUCTUATIONS OF THE INTEREST RATE ON 60 TO 90 DAY COMMERCIAL PAPER, NEW YORK CITY, 1903–1914

Deviations are plotted from the secular trend as zero. (Unit = one per cent. Data from Table 67, column (7), secular trend eliminated.)

TABLE 66. SECULAR TREND OF INTEREST RATES, 1903–1914

YEAR	DEVIATIONS FROM MID-POINT 1908–1909 (unit = half-year.)		INTEREST RATES (unit = 1 per cent)	DEVIATIONS FROM AVERAGE (4.89)	xy PRODUCTS (6)		SLOPE = −.02. Trend line values. Origin 1908–1909
X (1)	x (2)	x^2 (3)	Y (4)	y (5)	+	−	(7)
1903	−11	121	5.47	+ .58		6.38	5.00
1904	− 9	81	4.21	− .68	6.12		4.98
1905	− 7	49	4.40	− .49	3.43		4.96
1906	− 5	25	5.68	+ .79		3.95	4.94
1907	− 3	9	6.36	+1.47		4.41	4.92
1908	− 1	1	4.38	− .51	.51		4.90
Origin			**4.89**				**4.89**
1909	1	1	3.98	− .91		.91	4.88
1910	3	9	5.00	+ .11	.33		4.86
1911	5	25	4.03	− .86		4.30	4.84
1912	7	49	4.74	− .15		1.05	4.82
1913	9	81	5.60	+ .71	6.39		4.80
1914	11	121	4.78	− .11		1.21	4.78
		572	Average = 4.89		+16.78	−22.21	

$$\text{Slope of trend} = \frac{\Sigma xy}{\Sigma x^2} = \frac{\dfrac{+16.78 - 22.21}{2}}{572 \div 4} = -.02\%$$

This means that the *average annual change* in interest rates, 1903–1914, is .02 per cent in the *downward direction*. The reason for dividing the numerator by 2 and the denominator by 4 has been explained. The secular trend for this period is very slight, practically negligible.

Figures 44 and 45 present the results shown in Tables 65 and 66. The trend of interest rates moves in the opposite direction from that of pig-iron production but the trend is so slight in the former that it would introduce no serious error in comparing the short-time fluctuations of the two series if we did not eliminate the secular trend from the interest series. This can be demonstrated by correlating the two series, first with the secular trends eliminated from both, and then with the secular trend eliminated from the pig-iron series but not from the interest rate series. In the latter case the deviations in interest rates should be measured from the average for the 12-year period, 4.89 per cent, instead of measuring them from the annual secular trend values.

Figures 46 and 47 show the short-time fluctuations of the two series with the secular trends eliminated. The deviations of the original data

TABLE 67. CORRELATION OF THE FLUCTUATIONS OF PIG-IRON PRODUCTION AND INTEREST RATES, 1903–1914

(Secular trends eliminated from both series)

YEAR	ACTUAL PIG-IRON PRODUCTION (Unit = 1000 long tons)	TREND VALUES (Slope = 72)	DEVIATIONS OF (1) FROM (2)		ACTUAL INTEREST RATES (Unit = 1 per cent)	TREND VALUES (Slope = −.02)	DEVIATIONS OF (5) FROM (6)		*xy* PRODUCTS (9)	
	X (1)	(2)	x (3)	x^2 (4)	Y (5)	(6)	y (7)	y^2 (8)	+	−
1903	1452	1552	− 100	10,000	5.47	5.00	+ .47	.2209		47.00
1904	1344	1624	− 280	78,400	4.21	4.98	− .77	.5929	215.60	
1905	1882	1696	+ 186	34,596	4.40	4.96	− .56	.3136		104.16
1906	2066	1768	+ 298	88,804	5.68	4.94	+ .74	.5476	220.52	
1907	2109	1840	+ 269	72,361	6.36	4.92	+1.44	2.0736	387.36	
1908	1302	1912	− 610	372,100	4.38	4.90	− .52	.2704	317.20	
1909	2116	1984	+ 132	17,424	3.98	4.88	− .90	.8100		118.80
1910	2237	2056	+ 181	32,761	5.00	4.86	+ .14	.0196	25.34	
1911	1944	2128	− 184	33,856	4.03	4.84	− .81	.6561	149.04	
1912	2448	2200	+ 248	61,504	4.74	4.82	− .08	.0064		19.84
1913	2560	2272	+ 288	82,944	5.60	4.80	+ .80	.6400	230.40	
1914	1921	2344	− 423	178,929	4.78	4.78	.00	.0000		.00
			[a] −1597 +1602	1,063,679			[a] −3.64 +3.59	6.1511	+1545.46	−289.80
Average =	1948	....	....	....	4.89					

$$\sigma_X = \sqrt{\frac{1,063,679}{12}} = 298, \quad \sigma_Y = \sqrt{\frac{6.1511}{12}} = .72 \text{ per cent}$$

$$\text{and } r = \frac{+1545.46 - 289.80}{12 \text{ times } 298 \text{ times } .72} = +.49$$

[a] The plus and minus sums in these columns do not balance exactly due to the neglect of fractions in the computations. No correction factor is required, however, as in the case of the moving average.

from the annual values of the secular trend have been plotted, *showing cyclical movements*. The data are represented in the original units of tons and per cents, *which are not comparable.* A method of making the deviations comparable is presented later.

By comparing the correlation shown in Table 67 with that presented in Table 63 it will be observed that the coefficient is slightly higher when the secular trend has been eliminated by the straight line than when the moving average was used for the same purpose. However, in this particular problem of pig-iron production and interest rates, for the period 1903–1914, the coefficients obtained by the two methods differ only slightly (+.47 as compared with +.49).

Furthermore, since the secular trend of the interest rate series is slight for this particular period, the coefficient, r, is not affected appreciably if we do not take the trouble to eliminate the trend from the interest series before correlating it with pig-iron production. In Table 67, if the devia-

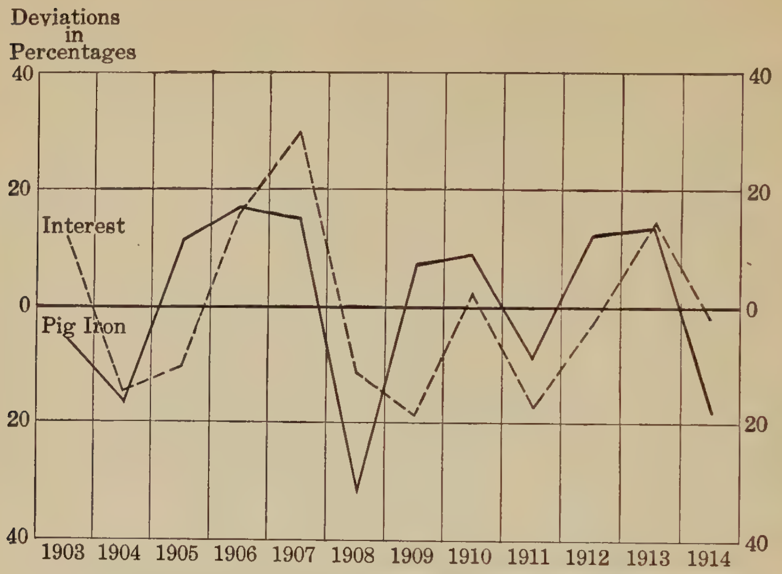


FIG. 48. COMPARISON OF THE CYCLICAL FLUCTUATIONS OF PIG-IRON PRODUCTION (CONTINUOUS LINE) WITH THOSE OF THE INTEREST RATE ON COMMERCIAL PAPER (DOTTED LINE), 1903–1914

(Data from Table 68, column (3) for Pig Iron, and column (7) for Interest Rates. The percentage deviations are taken from the trend values for Pig Iron, and from the average for 12 years, 1903–1914, for Interest Rates, because for the latter the trend is so small.)

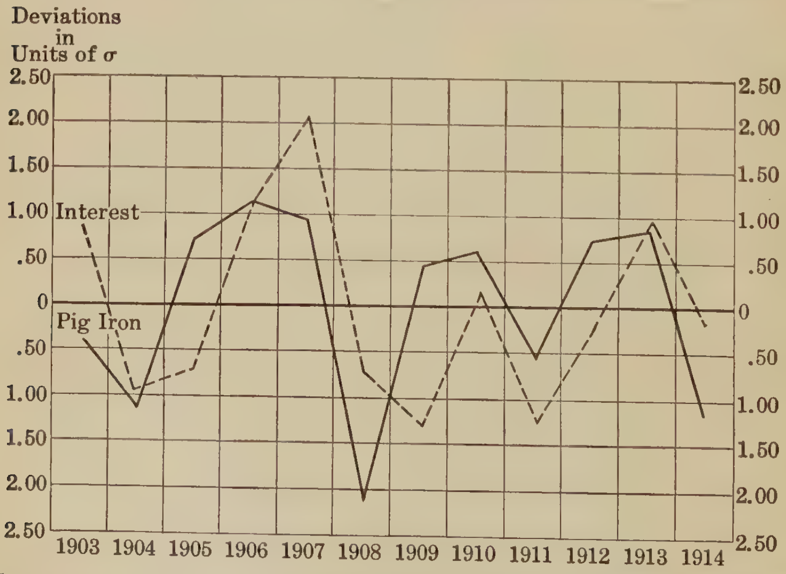


FIG. 49. COMPARISON OF THE CYCLICAL FLUCTUATIONS OF PIG-IRON PRODUCTION (CONTINUOUS LINE) AND OF THE INTEREST RATE (DOTTED LINE), 1903–1914

Deviations are expressed in units of the standard deviation of the percentages used in Fig. 48. (Data from Table 68, column (5) for Pig Iron, and column (9) for Interest Rates.)

TABLE 68. CORRELATION OF PERCENTAGE DEVIATIONS IN PIG-IRON PRODUCTION AND INTEREST RATES, 1903–1914

(Columns (5) and (9) express these deviations in Units of Standard Deviation, $\frac{x}{\sigma}$ and $\frac{y}{\sigma}$, and column (11) correlates them)

YEAR	PIG-IRON TREND	DEVIATIONS FROM TREND (Unit = 1000 tons)	PERCENTAGE DEVIATIONS FROM TREND (2) ÷ (1) × 100	SQUARES OF PERCENTAGES COLUMN (3)	DEVIATIONS IN σ UNITS OF PERCENTAGES (3) ÷ 15.2	INTEREST RATE DEVIATIONS FROM AVERAGE 4.89	PERCENTAGE DEVIATIONS FROM AVERAGE	SQUARES OF COLUMN (7)	DEVIATIONS IN σ UNITS OF PERCENTAGES (7) ÷ 14.7	xy PRODUCTS (3) × (7) (10)		PRODUCTS OF $\frac{x}{\sigma} \times \frac{y}{\sigma}$ (Units of σ) (5) × (9) (11)[c]	
	X (1)	x (2)	x (3)	x^2 (4)	$\frac{x}{\sigma}$ (5)	y (6)[a]	y (7)	y^2 (8)	$\frac{y}{\sigma}$ (9)	+	−	+	−
1903	1552	−100	− 6.4	40.96	− .42	+ .58	+11.9	141.61	+ .81		76.16		.3402
1904	1624	−280	−17.2	295.84	−1.13	− .68	−13.9	193.21	− .95	239.08		1.0735	
1905	1696	+186	+11.0	121.00	+ .72	− .49	−10.0	100.00	− .68		110.00		.4896
1906	1768	+298	+16.9	285.61	+1.11	+ .79	+16.2	262.44	+1.10	273.78		1.2210	
1907	1840	+269	+14.6	213.16	+ .96	+1.47	+30.1	906.01	+2.05	439.46		1.9680	
1908	1912	−610	−31.9	1017.61	−2.10	− .51	−10.4	108.16	− .71	331.76		1.4910	
1909	1984	+132	+ 6.7	44.89	+ .44	− .91	−18.6	345.96	−1.27		124.62		.5588
1910	2056	+181	+ 8.8	77.44	+ .58	+ .11	+ 2.2	4.84	+ .15	19.36		.0870	
1911	2128	−184	− 8.6	73.96	− .57	− .86	−17.6	309.76	−1.20	151.36		.6840	
1912	2200	+248	+11.3	127.69	+ .74	− .15	− 3.1	9.61	− .21		35.03		.1554
1913	2272	+288	+12.7	161.29	+ .84	+ .71	+14.5	210.25	+ .99	184.15		.8316	
1914	2344	−423	−18.0	324.00	−1.18	− .11	− 2.2	4.84	− .15	39.60		.1770	
		b	b	2783.45	b	b	b	2596.69	b	+1678.55	−345.81	+7.5331	−1.5440

$$\sigma_X \text{ Percentage Deviations} = \sqrt{\frac{2783.45}{12}} = 15.2 \text{ per cent}$$

$$\sigma_Y \text{ Percentage Deviations} = \sqrt{\frac{2596.69}{12}} = 14.7 \text{ per cent}$$

$$r = \frac{+1678.55 - 345.81}{12 \text{ times } 15.2 \text{ times } 14.7} = +.50$$

[a] It is assumed that there is *no secular trend in interest rates* because it is so slight. The deviations in column (6) are measured from the average of the original data for the 12 years, found in column (5) of Table 67 (4.89).

[b] The plus and minus values in these columns do not exactly balance because of the error in neglecting fractions at various points in the series of computations. The error is negligible.

[c] When the percentage deviations are expressed in units of standard deviation as in columns (5) and (9), it is easy to compute r by multiplying column (5) by column (9), taking the algebraic sum of the products in column (11), and dividing by 12, the number of pairs.

$$r = \frac{\Sigma\left(\frac{x}{\sigma} \text{ times } \frac{y}{\sigma}\right)}{12} = \frac{+7.5331 - 1.5440}{12} = +.50$$

which is identical to the second decimal place with the result obtained by correlating the percentage deviations of columns (3) and (7).

tions of the items in column (5) from the average for the twelve-year period (4.89) are entered in column (7) and these fluctuations are correlated with column (3) for pig iron, r is found to be +.49. This is the same value as that obtained when the secular trend of interest rates was eliminated. The student is requested to test this for himself.

DEVIATIONS FROM THE TREND EXPRESSED IN COMPARABLE UNITS — PERCENTAGES — STANDARD DEVIATIONS

In the comparison of the fluctuations of two or more series from their respective trends in terms of the original values we are likely to face the difficulty of incomparable units. One series is expressed in tons, the other in per cent, or some other unit of measurement. *The units may be made similar by setting forth the deviations from the trend as percentages of the trend line values* (Figure 48). This device removes the difficulty of comparing different units. All deviations are expressed as percentages of the trends and are plotted on similar scales.

But the difficulty in graphic presentation is not completely solved. Various series show different characteristics in their percentage deviations from the trends — *in the amplitude of their fluctuations.* How shall we compare the significance of the *amount of fluctuations* in two or more series? In all our correlations, both in this chapter and in the preceding one, we have made the two series comparable in this respect by setting forth the movements of each in terms of its own *standard deviation.* This unit forms a common measure of dispersion and measures the tendency of a given series to deviate from its own arithmetic average. The standard deviation may be computed for the deviations from the trend, expressed either in actual amounts or in percentages. In any case the scale for deviations from the trend will now be in units of σ (1σ, 2σ, 3σ, above or below the trend, as in Figure 49).

Table 68, column (3) expresses the actual deviations from the trend, column (2), as a percentage of the trend values for each year, column (1). For each year the value in (2) times 100 is divided by the value in (1). The values in column (6) are found by subtracting for each year the original data, found in column (5) of Table 67, from the average for the 12 years (4.89), on the assumption of *no secular trend in interest rates.* Then, for each year the values in column (6) times 100 are divided by 4.89 to obtain the percentage deviations of column (7). The percentage deviations of columns (3) and (7) are correlated by the usual method, resulting in $r = +.50$. This is practically identical with the coefficient obtained by correlating the short-time fluctuations expressed in tons and per cents of interest rates, as shown in Table 67.

Deviations in terms of standard deviation units. In the process of computing r we have found the standard deviations of the two series in terms of percentage fluctuations ($\sigma_X = 15.2$ per cent and $\sigma_Y = 14.7$ per cent). *These two σ's constitute a common unit for expressing in comparable terms the amount of the percentage fluctuations of the two series — their amplitude.* Column (5) values are obtained for each year by dividing the values of column (3) by 15.2, the σ_X. Likewise, the values of column (9) are obtained by dividing those of column (7) by 14.7, the σ_Y. Now the deviations are expressed in terms of σ and fractions of σ. In Table 68, column (11) and the footnote, the method of correlating the deviations expressed in units of the standard deviation is explained and illustrated. When each deviation is set forth in terms of the standard deviations of the series, the movements in the two series are directly comparable. The cyclical fluctuations are shown graphically both in terms of percentage deviations and in terms of units of standard deviations in Figures 48 and 49.

It should be noted that if the trend is determined by the method of a moving average, as was done in the early part of this chapter, the fluctuations about the moving average may be expressed in terms of percentages or units of standard deviation. Since the deviations plus and minus from the moving average do not usually balance, it is necessary to make a correction in computing the standard deviation, as in Table 63. In other respects the method is identical with the one just described.

It is possible to express the absolute fluctuations in terms of the standard deviation without reducing them first to percentages of the trend, as will be clear by reference to Table 67.

A COMPOSITE CURVE SHOWING SHORT-TIME FLUCTUATIONS

In the preceding pages methods have been explained for the description of the cyclical movements of a *single time series,* after testing for and eliminating, if present, the influence of long-time changes or secular trends. Correlation methods have been used to indicate the importance of eliminating trends while the cyclical movements of one series are being compared with those of other series, and for the purpose of measuring the relationship between the movements of fundamental series of economic facts. The description of a single series has finally taken the form of percentage deviations from the trend or the average, and the measurement of the amount of these deviations in units of standard deviation. The last step renders different series comparable from the point of view of their *amplitude.* Graphic methods have been emphasized as fundamental in observing the time, the nature and the degree of such fluctuations.

But, fundamental economic conditions are not shown adequately by a single series. It becomes necessary to combine single series, each treated by methods such as we have described, into a *composite curve.* This is a conclusive argument for describing each series in comparable units which may be combined with other series, assigning to each the proper weight, for any year or month. Such a composite curve may be used to describe the movement of business conditions in specific fields — *the business cycle.* This problem is similar in character to the construction of an index number. The publications of the Harvard Committee on Economic Research will furnish the student with abundant material in the application of these methods. In Figure 50 is shown a section of the composite curve of business conditions constructed by the American Telephone and Telegraph Company.[1]

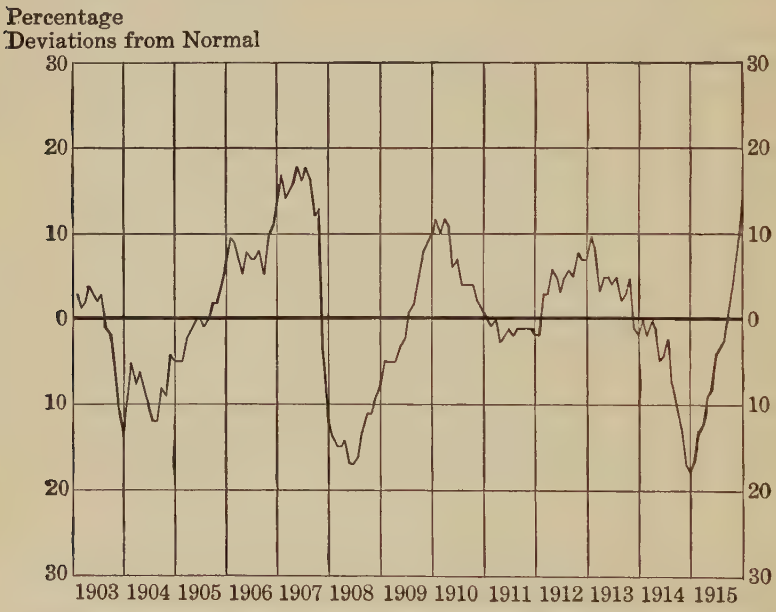


FIG. 50. COMPOSITE CURVE REPRESENTING THE MOVEMENT OF GENERAL BUSINESS CONDITIONS IN THE UNITED STATES, 1903–1915

(Data compiled and combined by the American Telephone and Telegraph Company, Office of the Chief Statistician. By permission of the Chief Statistician.)

[1] The use of such a curve is illustrated in an article by William F. Ogburn and Dorothy S. Thomas, "The Influence of the Business Cycle on Certain Social Conditions," *Quarterly Publication of the American Statistical Association*, September, 1922.

USE OF MOVING AVERAGES TO DESCRIBE TRENDS OTHER THAN STRAIGHT LINES

The moving average is frequently employed to describe the long-time changes when a straight line is clearly not the best fit to the original data, or when the direction of the trend changes.[1] The alternatives are to fit a more complicated curve, for example, a parabola of the second order; or to break up the entire series into two or more parts and to describe each part by its own straight line. This procedure renders more difficult the description of the entire series.

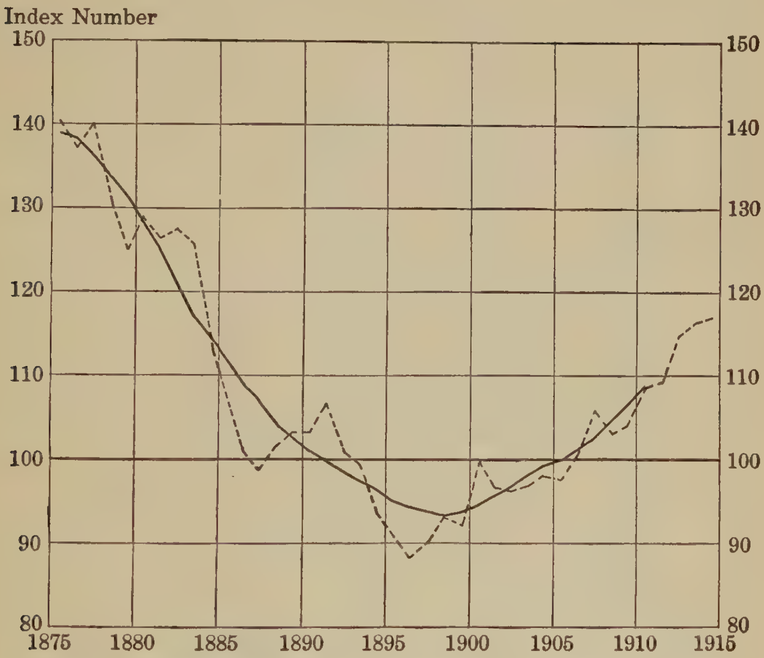


FIG. 51. MOVEMENT OF WHOLESALE PRICES AS REPRESENTED BY THE BRITISH BOARD OF TRADE INDEX NUMBERS, 1875–1914

The trend is determined by a nine-year moving average (the continuous line). (Data from Table 69, page 336. Dotted line represents the original annual index numbers.)

For illustration the index number of the Board of Trade of Great Britain for wholesale prices, 1871 to 1914, is presented in Table 69 and Figure 51. The trend is decidedly downward until about 1896, and after that date the movement is gradually upward until the opening of the

[1] The reader should refer to page 320 where the limitations of the moving average were stated. When the trend is parabolic or decidedly non-linear the moving average tends to change its form.

World War. A nine-year moving average is employed to describe this trend throughout the period.

TABLE 69. BOARD OF TRADE INDEX NUMBER, WHOLESALE PRICES, 1871–1914[a]

(Average prices 1900 = 100)

YEAR (1)	INDEX (2)	NINE-YEAR MOVING AVERAGE (3)	YEAR (1)	INDEX (2)	NINE-YEAR MOVING AVERAGE (3)
1871	135.6		1894	93.5	96.3
1872	145.2		1895	90.7	95.0
1873	151.9		1896	88.2	94.3
1874	146.9		1897	90.1	93.8
1875	140.4	139.3	1898	93.2	93.4
1876	137.1	138.6	1899	92.2	93.8
1877	140.4	136.5	1900	100.0	94.7
1878	131.1	133.8	1901	96.7	95.7
1879	125.0	131.5	1902	96.4	96.9
1880	129.0	128.5	1903	96.9	98.3
1881	126.6	125.2	1904	98.2	99.5
1882	127.7	120.8	1905	97.6	100.0
1883	125.9	117.2	1906	100.8	101.3
1884	114.1	114.7	1907	106.0	102.8
1885	107.0	111.8	1908	103.0	104.8
1886	101.0	109.2	1909	104.1	106.8
1887	98.8	106.9	1910	108.8	109.0
1888	101.8	104.2	1911	109.4	
1889	103.4	102.5	1912	114.9	
1890	103.3	101.0	1913	116.5	
1891	106.9	99.9	1914	117.2	
1892	101.1	98.7			
1893	99.4	97.4			

[a] Data from Bulletin 284, United States Bureau of Labor Statistics, Table 56, p. 265.

TESTING FOR LAG AND DETERMINING ITS AMOUNT

The meaning of lag has been explained. It remains to describe a method for detecting the presence of a lag when two series or combinations of series are being compared, and to measure the amount by which one series anticipates or falls behind another in its movements. When the production of pig iron fluctuates above its trend or normal does the change in interest rate occur at once, several months or a year later, or has this movement in interest rate preceded the other?

This fact becomes of fundamental importance in forecasting, provided the relationship between the movements of the two series or combination of series, as tested from relatively long past experience, is close. The size of r tests the closeness of relationship. Knowing from past experience the degree of correspondence between the cyclical movements and their direction (as shown by the correlation coefficient), and knowing the amount of time by which the one series anticipates or leads the other in its movement, we are able to predict values for the lagging series at specific points of time in the future, not with absolute exactness but within a range of error.

An important reason for recording economic and social data by months is now clear. Measurable lags, usually, do not occur in years. So far we have used only annual data for the sake of simplicity in presenting methods and in making computations. We shall test the two series *on the hypothesis of one year lag in interest rates,* knowing that it is probably not so great. The monthly data is analyzed in the final section of this chapter to determine the seasonal variations. Then it is possible to test the series for a lag shorter than one year, by correlating monthly movements.

A close inspection of Figure 49 suggests that the movement of interest rates follows that of pig-iron production and with fairly close correspondence in direction. The highest correlation coefficient obtained thus far, by relating the movements of identical years, is +.50 in Table 68. *But we can pair other than identical years. Will the degree of correspondence become closer* if we assume that the movement in the interest rate occurs one year later than that for pig iron? To test this assumption let us move the entire series of interest rates back one year, correlating the values for 1904–1915 with those for pig-iron production 1903–1914, in Table 70, page 338.

Column (4) of Table 70 gives the year following that stated in column (1) for pig iron. Therefore, the period covered by the interest rates is 1904–1915. Having obtained the average interest rate for this period and the deviations from it, column (6), *on the hypothesis of a negligible secular trend,* the procedure for correlating the short-time movements of the two series is already familiar. The resulting coefficient is +.65, as compared with +.50 (Table 68), which indicates a considerably closer association between the two movements when the one is assumed to occur a year later than the other. Would the association prove to be still closer (r be greater), if some period less than a year were assumed to measure the lag?

SEASONAL VARIATIONS — MONTHLY DATA

If the data are available by months it is possible to test the amount of lag by pairing the interest rate fluctuation for February with that for pig-iron production in January, for March with January, for April with January, and so on, assuming lags of one, two, three or more months in succession, until the highest value for r is obtained.

TABLE 70. CORRELATING THE CYCLES OF PIG-IRON PRODUCTION 1903–1914 WITH THE CYCLES OF INTEREST RATES 1904–1915 — LAG ONE YEAR

(Units of deviation in actual amounts — tons and per cent)

YEAR PIG-IRON SERIES	DEVIATIONS FROM TREND		YEAR INTEREST RATE SERIES	ACTUAL INTEREST RATES	DEVIATIONS FROM AVERAGE [a]		xy PRODUCTS (8)	
(1)	x (2)	x^2 (3)	(4)	Y (5)	y (6)	y^2 (7)	+	−
1903	−100	10,000	1904	4.21	− .51	.2601	51.00	
1904	−280	78,400	1905	4.40	− .32	.1024	89.60	
1905	+186	34,596	1906	5.68	+ .96	.9216	178.56	
1906	+298	88,804	1907	6.36	+1.64	2.6896	488.72	
1907	+269	72,361	1908	4.38	− .34	.1156		91.46
1908	−610	372,100	1909	3.98	− .74	.5476	451.40	
1909	+132	17,424	1910	5.00	+ .28	.0784	36.96	
1910	+181	32,761	1911	4.03	− .69	.4761		124.89
1911	−184	33,856	1912	4.74	+ .02	.0004		3.68
1912	+248	61,504	1913	5.60	+ .88	.7744	218.24	
1913	+288	82,944	1914	4.78	+ .06	.0036	17.28	
1914	−423	178,929	1915	3.45	−1.27	1.6129	537.21	
	[b]	1,063,679			[b]	7.5827	+2068.97	−220.03
Average......	....		....	4.72				

$$\sigma_X = \sqrt{\frac{1{,}063{,}679}{12}} = 298$$

$$\sigma_Y = \sqrt{\frac{7.5827}{12}} = .79 \text{ per cent}$$

$$r = \frac{+2068.97 - 220.03}{12 \times 298 \times .79} = +.65$$

[a] It is assumed that the secular trend is negligible for interest rates 1904–1915, as in Table 68.

[b] Differences between plus and minus sums are negligible, being due to fractions.

So far the time series has been treated so as to eliminate the influence of secular trend while the cyclical movements were being examined. As soon as monthly figures are secured another type of fluctuation is introduced for certain kinds of data — *the seasonal.* Employment, production, prices, vary within the year, as shown in Figure 52. If we attempt to do what was suggested in the last paragraph, in testing the amount of lag, *seasonal variations will be confused with the cyclical movements which are being analyzed and examined.* The seasonal movement may be downward at the time the cyclical movement is in the opposite direction. For example, employment may be relatively low in December while the general cycle of prosperity is moving steadily upward over a period of several years. At other times of the year the two movements reinforce each other. Therefore, *if we mean to use monthly data, some* *method is required for the purpose of eliminating seasonal fluctuations as well as secular trends while the cyclical movements are being investigated.*

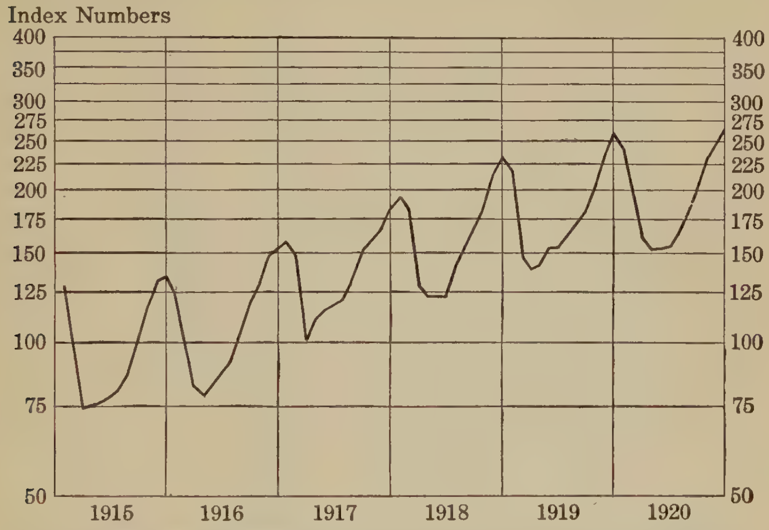


FIG. 52. SEASONAL MOVEMENTS IN THE RETAIL PRICES OF EGGS, 1915–1920

(Average prices for the United States in 1913 = 100. Data from Bulletin 315, Bureau of Labor Statistics, *Retail Prices*, Table A, pp. 87 and 89. See diagram in this Bulletin, p. 29. The ratio vertical scale is used as in the Bulletin.)

AN INDEX OF SEASONAL VARIATION

For many purposes data recorded for shorter periods than one year are essential. Without monthly or quarterly figures we are frequently unable to measure the amount of lag satisfactorily. When weekly and monthly data are used in the analysis of cyclical movements, it is desirable to ascertain whether there exists a well defined seasonal variation, and, if so, to obtain an index of its amount at any given month in order to eliminate this disturbing factor in the study of cycles.

Different methods are employed in constructing this index. The explanation of some of these will furnish the student with valuable illustrations of the applications of statistical methods. We shall not undertake a critical and final evaluation of the various procedures employed, but shall attempt to emphasize the nature of the problem of the measurement of seasonal variation. *The problem is to generalize from experience a typical seasonal movement and to express this in monthly indexes which may be applied to the original data for each year with the purpose of eliminating the seasonal influence.*

I. Method of simple monthly means. In Table 71 the production of pig iron for all Januaries, 1903–1914, is averaged to obtain a typical amount for this month; and the same procedure is followed for each successive month of the year.

TABLE 71. PIG-IRON PRODUCTION — MONTHLY TYPES 1903–1914 [a]

(Unit = 1000 long tons)

YEAR	JAN.	FEB.	MAR.	APR.	MAY	JUNE	JULY	AUG.	SEPT.	OCT.	NOV.	DEC.	ANNUAL AVERAGE
1903	1472	1390	1590	1608	1713	1673	1546	1571	1553	1425	1039	846	1452
1904	921	1205	1447	1555	1534	1292	1106	1167	1352	1450	1486	1616	1344
1905	1781	1597	1936	1922	1963	1793	1741	1843	1899	2053	2014	2045	1882
1906	2068	1904	2155	2073	2098	1976	2013	1926	1960	2196	2187	2235	2066
1907	2205	2045	2226	2216	2295	2234	2255	2250	2183	2336	1828	1234	2109
1908	1045	1077	1228	1149	1165	1092	1218	1348	1418	1563	1577	1740	1302
1909	1801	1703	1832	1738	1880	1929	2101	2246	2385	2600	2547	2635	2116
1910	2608	2397	2617	2483	2390	2265	2148	2106	2056	2093	1909	1777	2237
1911	1759	1794	2188	2065	1893	1787	1793	1926	1977	2102	1999	2043	1944
1912	2057	2100	2405	2375	2512	2440	2410	2512	2463	2689	2630	2782	2448
1913	2795	2586	2763	2752	2822	2628	2560	2543	2505	2546	2233	1983	2560
1914	1885	1888	2348	2270	2093	1918	1958	1995	1883	1778	1518	1516	1921
Monthly Types (A)	22397 1866	21686 1807	24735 2061	24206 2017	24358 2030	23027 1919	22849 1904	23433 1953	23634 1970	24831 2069	22967 1914	22452 1871	23381 1948
Seasonal Index (B) 1948 = 100	95.8	92.8	105.8	103.5	104.2	98.5	97.7	100.3	101.1	106.2	98.3	96.0	100.0

[a] Data from *Review of Economic Statistics*, Preliminary Volume I, p. 66, Committee on Economic Research, Harvard University.

The monthly types of row (A) at the bottom of the table are each reduced to a percentage of the average monthly production for the entire period, 1948 unit-tons = 100, by dividing each of the values in row (A) by 1948 and multiplying by 100. The resulting percentages in row (B) furnish indexes of seasonal fluctuations, generalized from the experience of 12 years.

The index in this simple form does not make allowance for the secular trend. Professor George R. Davies, in his *Introduction to Economic Statistics*, pp. 116–20, uses this method of monthly means to obtain a seasonal index for interest rates, by relating the monthly types such as the values in row (A) of Table 71 to the corresponding monthly trend values for the middle twelve months of the entire period. This method makes allowance for the secular trend which characterizes any year of the twelve-year period. Table 72 illustrates the method for pig-iron production, 1903–1914.

The mean monthly production for the period 1903–1914 is 1948 unit-tons. This amount would be plotted between December, 1908 and January, 1909, at the middle of the period, because the number of years (12) is even. The annual secular trend has been determined already in Table 65, showing a yearly increment of 72 unit-tons (slope = 72). The trend

TABLE 72. MONTHLY TYPES RELATED TO SECULAR TREND

Middle 12 months of period 1903–1914, July, 1908–June, 1909 (1)	Averages for each month from Table 71, row (A) (2)	Monthly trend for middle 12 months (3)	Seasonal index (2) ÷ (3) × 100 (4)	Arranging months Jan. to Dec. (5)	Jan. to Dec. indexes by months. Data from column (4) (6)	Monthly deviation from normal = 100 (7)
1908 July	1904	1915	99.4	Jan.	95.6	−4.4
Aug.	1953	1921	101.7	Feb.	92.3	−7.7
Sept.	1970	1927	102.2	March	105.0	+5.0
Oct.	2069	1933	107.0	April	102.4	+2.4
Nov.	1914	1939	98.7	May	102.8	+2.8
Dec.	1871	1945	96.2	June	96.9	−3.1
Middle of period 1903–1914		1948				
Jan.	1866	1951	95.6	July	99.4	−0.6
Feb.	1807	1957	92.3	Aug.	101.7	+1.7
March	2061	1963	105.0	Sept.	102.2	+2.2
April	2017	1969	102.4	Oct.	107.0	+7.0
May	2030	1975	102.8	Nov.	98.7	−1.3
1909 June	1919	1981	96.9	Dec.	96.2	−3.8
			100.0		100.0	

line was fitted by the method of least squares to the annual averages. Dividing the annual increment by 12 gives the monthly increase due to the growth factor, 6 unit-tons. One half of this amount, 3 unit-tons, must be subtracted from 1948 to obtain the value of the ordinate in the middle of December, 1908; likewise, 3 unit-tons must be added to 1948 to obtain the value for January, 1909. The values for the other months below or above the mid-point of column (3) are obtained by adding or subtracting 6 unit-tons for each month of change in time. These monthly trend values in column (3), for the middle twelve months of the period made up of the last six months of 1908 and the first six months of 1909, may be regarded as the type trend values for the entire period, since the slope of the trend line is uniform for each year. Since these trend values are located at the middle of the period, the monthly types of Table 71, row (A), may be related to them in the form of percentages for each corresponding month.

The influence of the secular trend upon the seasonal index is eliminated by dividing the monthly types of column (2) by the trend values of column (3), item by item, instead of using 1948 unit-tons as the constant divisor, as was done in Table 71. If these quotients in column (4) do not average 100 (as they do in this case) they must be adjusted by dividing each percentage by the average of the items in column (4).

Columns (5) and (6) merely rearrange the monthly seasonal indexes of column (4) from January to December for convenience, and column (7) describes the deviations of each month from the normal (100) for the year. Since there are several cycles in the period under consideration it is assumed that their effect upon the seasonal index is canceled by taking monthly means of all Januaries, Februaries, etc. for the entire period 1903–1914. This emphasizes the importance of taking a fairly long period of years in constructing a seasonal index.[1]

The *generalized index* shown in column (6) may be applied to each year of the twelve-year period for the purpose of eliminating the factor of seasonal variation while the cycles are being examined.

A fundamental defect in the method just described is that *the monthly means may not be typical because of the influence of extreme and irregular fluctuations upon the average for any specific month.* In practice this method proves to be untrustworthy. This defect could be avoided by arraying all January values in Table 71 in order of size and adopting the *median* value or an average of two or more central values in the array as typical of the given month. The same procedure could be followed for each month and the monthly types thus obtained could be expressed as percentages of the trend values shown in Table 72, column (3). By this method the influence of extreme variants upon the monthly types would be eliminated.

A simple alternative method for avoiding the influence of extreme variants would be to express the original values as percentages of the monthly trend values for each month in the entire period of 12 years, as in column (6), Table 78. All January percentages could be arrayed in order of size, and the median or an average of two or more of the central values could be taken as typical of that month. All the other months

[1] Objection may be raised to this use of averages on the ground that the seasonal movement changes in character over a period of years for the given type of data. See Willford I. King, "An Improved Method for Measuring the Seasonal Factor," *Journal of the American Statistical Association*, September, 1924, for the presentation of a method designed to meet this objection. Also W. L. Crum, " Progressive Variation in Seasonality," *Journal of the American Statistical Association*, March, 1925.

could be treated in the same manner. Then the monthly types obtained in this manner could be adjusted to average 100 for the twelve months. This method of obtaining a seasonal index will be referred to later, under IV.

II. Method of a twelve-month moving average. A simple and direct method of smoothing out seasonal fluctuations consists in obtaining a twelve-month moving average for the period. The values thus computed fall at the first of each month instead of the middle, as in the original series, for example the average of the twelve items, January to December falls between June and July; the next average falls between July

Table 73. Twelve-Month Moving Averages of Pig-Iron Production 1903–1914

(Light-faced type figures located at first of each month because of even number of months. Bold-faced type figures centered at middle of month by two-month average.)

Month	1903	1904	1905	1906	1907	1908	1909	1910	1911	1912	1913	1914
January . . .	[a] 1517 **1519**	1328 **1310**	1597 **1623**	1989 **2000**	2145 **2155**	1570 **1527**	1646 **1683**	2440 **2442**	1965 **1950**	2144 **2170**	2653 **2659**	2231 **2206**
February . .	1521 **1526**	1291 **1274**	1650 **1679**	2012 **2016**	2165 **2178**	1484 **1447**	1719 **1756**	2443 **2437**	1935 **1927**	2196 **2220**	2665 **2666**	2181 **2158**
March.	1532 **1537**	1258 **1250**	1707 **1729**	2019 **2021**	2192 **2202**	1409 **1377**	1794 **1835**	2432 **2418**	1920 **1917**	2244 **2265**	2668 **2670**	2135 **2109**
April	1542 **1540**	1241 **1242**	1752 **1778**	2024 **2030**	2211 **2216**	1345 **1312**	1875 **1918**	2404 **2383**	1913 **1913**	2285 **2309**	2671 **2665**	2083 **2051**
May.	1538 **1521**	1243 **1261**	1803 **1825**	2036 **2043**	2222 **2207**	1280 **1270**	1961 **2001**	2362 **2336**	1914 **1918**	2334 **2360**	2659 **2642**	2019 **1989**
June.	1504 **1478**	1280 **1312**	1847 **1864**	2050 **2058**	2192 **2151**	1260 **1281**	2042 **2079**	2309 **2273**	1922 **1933**	2386 **2417**	2626 **2593**	1960 **1941**
July.	1452 **1429**	1344 **1380**	1882 **1894**	2066 **2072**	2109 **2060**	1302 **1334**	2116 **2150**	2237 **2202**	1944 **1957**	2448 **2479**	2560 **2522**	1921 **1909**
August	1406 **1399**	1416 **1433**	1906 **1919**	2077 **2083**	2012 **1972**	1365 **1391**	2184 **2213**	2167 **2141**	1969 **1981**	2509 **2529**	2484 **2455**	1897 **1888**
September.	1391 **1385**	1449 **1469**	1932 **1941**	2089 **2092**	1932 **1890**	1417 **1442**	2242 **2275**	2116 **2099**	1994 **2003**	2550 **2565**	2426 **2409**	1879 **1867**
October. . . .	1379 **1377**	1489 **1504**	1950 **1957**	2095 **2101**	1848 **1804**	1467 **1491**	2307 **2338**	2081 **2063**	2012 **2025**	2580 **2596**	2391 **2371**	1856 **1850**
November .	1375 **1367**	1520 **1538**	1963 **1968**	2107 **2115**	1760 **1713**	1516 **1546**	2369 **2390**	2046 **2025**	2038 **2064**	2611 **2624**	2351 **2320**	1843 **1850**
December. .	1360 **1344**	1556 **1577**	1974 **1982**	2123 **2134**	1665 **1617**	1576 **1611**	2412 **2426**	2004 **1985**	2090 **2117**	2637 **2645**	2290 **2261**	1857 **1876** 1896

[a] The original data for the computation of these twelve-month averages in light-face type are found in Table 71, with the exception of the last six months of 1902 which are needed to compute the averages for the first six months of 1903. Likewise, the first six months of 1915 are required to compute averages for the last six months of 1914. The additional data are given below:

1902						1915					
July	Aug.	Sept.	Oct.	Nov.	Dec.	Jan.	Feb.	Mar.	Apr.	May	June
1505	1436	1432	1475	1440	1471	1601	1675	2064	2116	2263	2381

The monthly data for 1902 are obtained from "Monthly Production of Pig Iron 1884 to 1903," Margaret G. Myers, *Journal of the American Statistical Association*, June, 1922, p. 249; and the data for 1915 are taken from the same source as those in Table 71.

and August, etc. These values can easily be centered opposite the original items at the middle of each month by a derived two-month moving average. *In this manner the original item for each month is replaced by the twelve-month average.* Table 73 gives both the twelve-month moving averages and the two-month derived averages, the latter being printed in bold-face type.

The bold-face type figures in Table 73 are centered at the middle of each successive month by computing a derived two-month moving average from the figures in light-face type, which represent production values at the beginning of each month obtained by averaging an even number of months each time (12). For example, for January, 1903,

$$\frac{1517 + 1521}{2} = 1519, \text{ at middle of January}$$

$$\text{Likewise, } \frac{1521 + 1532}{2} = 1526\tfrac{1}{2} \text{ for February}$$

In any two-month average the result is always an integer or an integer and a fraction, ½. If the ½ is counted at the next higher integer, as is customary, then all errors are in the same direction, as would be the case also if the fraction were always dropped. In the *bold-face type figures* the fraction ½ has been counted at the next higher integer and dropped entirely *at alternate occurrences* in order to balance the errors.

The seasonal fluctuations are smoothed out by the twelve-month averages which, when related to the secular trend values, measure the cyclical movements about the trend, as shown in Figure 53.

In this simple form the method is not satisfactory, because it assumes that any twelve-month period of the entire series is as good as any other in representing a seasonal movement. *A generalized seasonal swing can be described.*

A method based upon the moving averages. The centered monthly moving averages described above are regarded as 100 and the percentage of each original monthly item to the corresponding twelve-month average is obtained. (Table 78, p. 354, column (2) ÷ (3) × 100 = Column (4).) The percentages in column (4) of Table 78 are now arrayed for each month in Table 74. The columns show the percentages for each month in order of size for the entire period of twelve years. A *monthly type* is obtained by the method of *medians*, which avoids the influence of extreme variants, a defect inherent in the method of monthly means first described. *These monthly medians measure the seasonal movement, above or below the normal represented by the twelve-month moving average.*

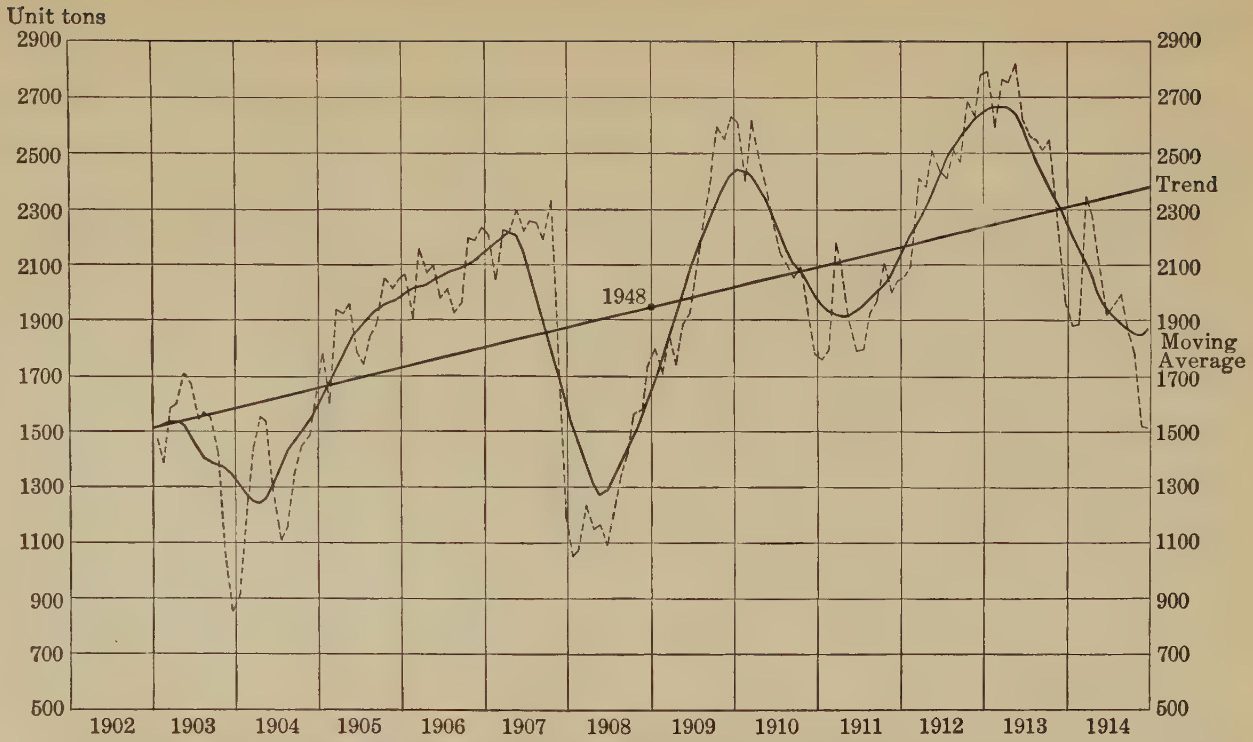


FIG. 53. MONTHLY PRODUCTION OF PIG IRON IN THE UNITED STATES (DOTTED LINE) AND THE MONTHLY SECULAR TREND (STRAIGHT LINE) 1903–1914

The monthly data are smoothed by a twelve-month moving average, centered at the middle of each month (continuous line showing cyclical movement about the trend after seasonal changes are smoothed). (Unit = 1000 long tons. Data from Table 78, p. 354, columns (2), (3) and (5). See footnote, Fig. 36.)

Table 74. Generalized Seasonal Index from Twelve-Month Moving Averages — 1903–1914

(Arrays of percentages for each month from column (4) Table 78, p. 354. Values for each month arranged from lowest to highest for twelve-year period.)

	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.	
	68.4	74.4	89.2	87.6	91.7	85.2	80.1	81.4	92.0	96.1	76.0	62.9	
	70.3	87.5	99.8	90.6	94.0	92.4	91.3	92.5	93.7	96.4	82.1	76.3	
	85.4	91.1	101.1	100.0	98.7	92.8	91.6	96.0	96.0	101.5	94.3	80.8	
	90.2	93.1	103.4	102.1	102.3	96.0	91.9	96.9	97.8	103.5	96.3	87.7	
	94.8	93.9	103.5	102.9	102.7	96.2	97.2	97.2	98.0	103.6	96.6	89.5	
[a]	96.9	94.4	106.2	103.3	104.0	98.5	97.2	98.4	98.3	103.8	96.9	96.5	[a]
	102.3	94.6	106.6	104.2	105.2	98.8	97.5	99.3	98.7	104.5	100 2	102.5	
	103.4	94.6	108.2	104.4	106.4	99.6	97.7	101.5	100.9	104.8	102.0	103.2	
	105.1	95.1	111.3	107.9	106.8	101.0	101.5	103.6	104.0	104.9	102.3	104.7	
	106.8	97.0	112.0	108.1	107.6	101.3	102.6	105.7	104.8	107.4	103.4	105.2	
	107.0	97.0	114.1	110.7	112.6	103.9	108.2	112.3	112.1	111.2	106.6	108.0	
	109.7	98.4	115.8	125.2	121.6	113.2	109.5	114.1	115.5	129.5	106.7	108.6	
Median values (A)	99.6	94.5	106.4	103.8	104.6	98.7	97.4	98.9	98.5	104.2	98.6	99.5	Ave. 100.4
Seasonal index (B)	99.2	94.1	106.0	103.4	104.2	98.3	97.0	98.5	98.1	103.8	98.2	99.1	Ave. 100.0
Deviation from 100 (C)	−.8	−5.9	+6.0	+3.4	+4.2	−1.7	−3.0	−1.5	−1.9	+3.8	−1.8	−.9	

[a] The *median value* is used to average the columns for the monthly types. Since the number of years employed, 1903–1914, is even (12 years), the twelve percentages are arrayed in each column for a specific month and the two mid-values are averaged to obtain the *median value*. This method avoids the influence of extreme fluctuations.

In Table 74 row (A) does not average 100 for the 12 months, but 100.4. Therefore, row (A) is adjusted to average exactly 100, by dividing each percentage in succession by 100.4 and multiplying by 100. The resulting percentage for each month is entered in row (B) and constitutes the final seasonal index for that month. This adjustment makes the percentages of row (B) average 100 for a normal year, obtained from the experience of the twelve-year period, 1903–1914. The deviations above and below 100 measure the seasonal fluctuations about the normal, and are entered in row (C).

This method, essentially as presented, has been employed by the Federal Reserve Bank of New York, and is referred to in Jordan's *Business Forecasting*, p. 212 (Prentice-Hall, 1923). It makes allowance for secular trend by the use of the moving average, especially when the trend is linear; and it is assumed that the cyclical influence is cancelled by the median monthly types. *By the use of the median* the distortion caused by extreme variants is avoided.

By experiment, when the number of years in the period analyzed is small, and when there is no marked tendency for the percentages in the columns of Table 74 to concentrate about the mid-value, *more typical values for the monthly indexes may be obtained by averaging several of the central values in each column, for example four or even six percentages. The problem is to isolate a typical seasonal movement from the fluctuations* which occur within each year over the period — fluctuations which are more or less irregular and dissimilar for the different years.

III. Method of link relatives.[1] This method has been developed and used by the Harvard Committee on Economic Research under the direction of Warren M. Persons. The steps in the computation of the index of seasonal movement may be stated as follows:

(1) Compute monthly link relatives for each month of the period, using the original data of the preceding month as the base for the link relative of the given month, $\frac{\text{Jan.}}{\text{Dec.}}$; $\frac{\text{Feb.}}{\text{Jan.}}$; $\frac{\text{Mar.}}{\text{Feb.}}$; $\frac{\text{Apr.}}{\text{Mar.}}$, etc.

(2) Array all January link relatives obtained from the entire period in order from lowest to highest values. Do the same for each month of the year.

(3) Take the *median relative* as the monthly type, thus avoiding the influence of extreme variants.

(4) Express each monthly type as a percentage based upon January (100), by multiplying or chaining progressively the median relatives from January to January. The final January product obtained by multiplying the December relative by the January relative usually does not equal 100 because of the influence of the secular trend in the process of chaining the median relatives.

(5) Distribute the difference between the computed January chain relative and 100 among all the monthly relatives of the chain series, with the object of making the computed January relative equal to 100. Use the arithmetic or the geometric principle of distributing this difference.

(6) Adjust the revised chain relatives so that their mean is equal to 100, by dividing by the average of the 12 chain relatives. *The resulting values for the successive months will be the monthly indexes of the typical seasonal movement.*

Table 75 shows link relatives for each month of the period 1903–1914, computed as explained by dividing the amount of pig-iron production of the given month by that of the preceding month and multiplying by 100.

[1] A link relative is the percentage which each item bears to the item of the preceding month, $\frac{\text{Jan.}}{\text{Dec.}} \times 100$, etc.

This procedure of linking tends to minimize the influence of cyclical changes upon the seasonal index. The influence of cycles is further reduced by taking a period long enough to include a number of cycles and in this manner to obtain typical median values for seasonal changes.

TABLE 75. LINK RELATIVES — PIG-IRON PRODUCTION 1903–1914

(Original data found in column (2), Table 78, p. 354.)

YEAR	JAN.	FEB.	MAR.	APR.	MAY	JUNE	JULY	AUG.	SEPT.	OCT.	NOV.	DEC.
	DEC.	JAN.	FEB.	MAR.	APR.	MAY	JUNE	JULY	AUG.	SEPT.	OCT.	NOV.
1903	(*a*)	94	114	101	*106	98	92	102	99	92	73	81
1904	109	131	120	107	99	84	86	106	116	107	102	109
1905	110	90	121	99	102	91	97	106	103	108	98	102
1906	101	92	113	96	101	94	102	96	102	112	100	102
1907	99	* 92	109	100	104	97	101	100	97	107	78	* 67
1908	85	103	114	94	101	94	112	*112	*104	*111	101	110
1909	103	95	*107	95	108	103	109	107	106	109	98	103
1910	99	92	109	95	96	95	95	98	98	102	91	93
1911	99	*101	122	94	92	* 95	100	107	103	106	95	102
1912	101	102	*114	99	106	97	99	104	98	109	98	106
1913	100	* 92	107	100	103	93	97	*100	99	102	88	89
1914	95	100	124	97	92	92	102	102	94	94	85	100

(*a*) Data on monthly production of pig iron in column (2) Table 78 are taken from the *Review of Economic Statistics*, Preliminary Volume I, p. 66, Committee on Economic Research, Harvard University. *The link relatives given above are found on p. 67 of the same volume.* Monthly quotations begin with January 1903. To compute a link relative for January 1903 it is necessary to estimate the production for December 1902. This estimate is found in the *Journal of the American Statistical Association*, June 1922, p. 249. The estimated amount for December 1902 is 1471 unit-tons. Then the link relative for January 1903 would be $\frac{1472}{1471} \times 100 = 100$.

* If the student computes the link relatives from the original production figures in Table 78, *by hand or by machine*, he will find slight discrepancies in the figures marked (*) in Table 75. *This is due to the fact that the Harvard Committee staff used a slide rule in the computation, which is not accurate to the same degree, in the second decimal place and following.* The author has used the Harvard Committee figures for the link relatives as given on page 67 of their volume because he wishes to encourage the student to make use of this very important source of monthly data for many different time series. Only one of the discrepancies noted (*), that for June, 1911 (95 which should be 94), affects in any way the computation of the seasonal index in Table 76. By substituting 94 the median link relative for June, in Table 76, would become 94 instead of 94.5. *The final effect upon the seasonal index is negligible.* In each case marked (*) the figure, if accurate to the nearest per cent, would read either one per cent less or one per cent more than that given in the table.

Median link relatives in row (*A*), Table 76. Table 76 in the January column ranks all the January relatives, taken from the $\frac{\text{Jan.}}{\text{Dec.}}$ column of Table 75, from lowest to highest values. A similar array of link relatives is shown in successive columns for each month. From these arrays it is possible to estimate the degree of regularity of the month-to-month (seasonal) changes during the twelve-year period 1903–1914. For any given month the closer the concentration of relatives about any value, the more significant and typical is the seasonal movement for that month. The columns show marked differences in the tendency to concentrate.

Since the number of items in each column is even (12), the median link relative for the January column is obtained by averaging the two

TABLE 76. ARRAYS OF LINK RELATIVES FOR EACH MONTH IN ORDER OF SIZE — SEASONAL INDEX

(Medians as Monthly Types. Relatives from Table 75 for each month.)

	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.	
	85	90	107	94	92	84	86	96	94	92	73	67	
	95	92	107	94	92	91	92	98	97	94	78	81	
	99	92	109	95	96	92	95	100	98	102	85	89	
	99	92	109	95	99	93	97	100	98	102	88	93	
	99	92	113	96	101	94	97	102	99	106	91	100	
Median	[a] (100)	94	114	97	101	94	99	102	99	107	95	102	
	100	95	114	99	102	95	100	104	102	107	98	102	
	101	100	114	99	103	95	101	106	103	108	98	102	
	101	101	120	100	104	97	102	106	103	109	98	103	
	103	102	121	100	106	97	102	107	104	109	100	106	
	109	103	122	101	106	98	109	107	106	111	101	109	
	110	131	124	107	108	103	112	112	116	112	102	110	Jan.
(A) Median link relatives	100.0	94.5	114.0	98.0	101.5	94.5	99.5	103.0	100.5	107.0	96.5	102.0	100.0 (A)
(B) Chain relatives to Jan. = 100.	100.0	94.5	107.7	105.5	107.1	101.2	100.7	103.7	104.2	111.5	107.6	109.8	109.8 (B)
Adjusted [b] (C) for trend	100.0	93.7	106.1	103.1	103.8	97.1	95.8	98.0	97.7	104.2	99.4	100.8	100.0 (C)
Seasonal (D) index	100.0	93.7	106.1	103.1	103.8	97.1	95.8	98.0	97.7	104.2	99.4	100.8	(D)

[a] If we use estimated production for December 1902 (1471 unit-tons) the January link relative for 1903 becomes $\frac{1472}{1471} \times 100 = 100$ (see footnote to Table 75). The median of the January column = 100.

If we omit the January link relative for 1903 altogether, then there are eleven items in the January column of this table and the median is 100, the same as before.

[b] We may use either the *arithmetic* or *geometric* principle of adjustment to make the final January chain relative = 100. We used *arithmetic principle* in row (C), that is, 1/12 of 9.8; 2/12 of 9.8; etc., subtracted from successive chain relatives in row (B). See page 350 in text.

middle items $\left(\frac{100 + 100}{2} = 100\right)$. All the other relatives in row (A) are obtained in the same manner. If the number of items in each column were odd the median would be the middle value in the array. *The resulting twelve median link relatives measure the relation of each month to the preceding month, each monthly value being a type for the twelve-year period.*

Chain relatives in row (*B*). The next step is to relate each relative in row (*A*) to the January relative as a common base (100). The resulting relative may be called a chain relative of the given month to January as a base, and is entered in row (*B*). The February chain relative is identical with the median link relative (94.5 × 100.0%). The March chain relative is the product of the March median link relative by the February chain relative (114.0 × 94.5% = 107.7%). In the same manner each median link relative is multiplied by the chain relative for the preceding month, throughout the year, month-by-month. Finally, a new January chain relative is obtained from the product of the January median link relative by the December chain relative (100.0 × 109.8% = 109.8%). This final product, *the new January relative,* should equal the original January relative, 100%, *provided other than seasonal influences do not disturb the result.* Actually, the difference is 9.8% (109.8 − 100.0 = 9.8). This discrepancy is *due mainly to the presence of secular trend in pig-iron production.*

Adjustment of chain relatives in row (C). The difference of 9.8%, due to other than seasonal influences, must be eliminated by deducting a proportion of it from each of the monthly chain relatives in succession from February to the following January. The object of this adjustment is to make the final January relative (100%) identical with the original January relative. In other words we wish to eliminate the influence of secular trend from the seasonal indexes.

Each of the median link relatives contains a small error due mainly to the presence of *secular trend.* The chain relatives are products of the median link relatives and the errors cumulate through the twelve products. Two methods of distributing this total error, 9.8%, are possible, the *arithmetic* and the *geometric.*[1] In most cases the two methods produce results without significant differences. The arithmetic method is simpler and has been employed in row (C) of the table, as illustrated:

94.5 − (1/12 of 9.8) = 93.7, the February adjusted relative
107.7 − (2/12 of 9.8) = 106.1, the March adjusted relative
........etc.............etc.............
109.8 − (11/12 of 9.8) = 100.8, the December adjusted relative
109.8 − (12/12 of 9.8) = 100.0, the final January relative.

The seasonal indexes in row (D). The adjusted chain relatives of row (*C*) are all percentages of the January base (100). *The final step is to make the base the average for a normal year* and to reduce the values in row (*C*) to percentages of this base (100). To do this it is necessary merely to average the twelve items in row (*C*), 99.98, and to divide each month's adjusted chain relative by this average and multiply by 100,

[1] The geometric principle of distributing the error is probably more defensible in theory. Let the constant error in each median link relative in row (A) be represented by d. Applying the constant factor $(1 + d)$ to the monthly medians, the final computed January chain relative equals $100\,(1 + d)^{12} = 109.8$, from which the value of $(1 + d)$ can be computed readily by the use of logarithms. Now the successive chain relatives in row (*B*), beginning with February, can be adjusted by dividing them by $(1 + d)$, $(1 + d)^2$, $(1 + d)^3 \ldots . (1 + d)^{11}$, $(1 + d)^{12}$. Dividing the computed January chain relative by $(1 + d)^{12}$ causes the entire error to disappear and produces the original index of 100 per cent. See "Correlation of Time Series," Warren M. Persons, *Journal of the American Statistical Association*, June, 1923, pp. 716–717, or *Handbook of Mathematical Statistics*, pp. 152–154.

$\frac{100.0}{99.98}$, $\frac{93.7}{99.98}$ times 100, etc. The resulting percentage for each of the twelve months of the normal year is entered in row (D).

The average of the twelve items in row (C) *in this particular problem* happens to be so near 100.0% (99.98) that when we divide each of the values in (C) by the average, 99.98, all of the percentages remain unchanged. Therefore, row (D) is identical with row (C) in the table, *although this would not be the case as a rule. The percentages in row* (D) *are the final indexes of seasonal variation, generalized from the experience of a twelve-year period.*[1]

A special merit of the link relative method is that it permits us to array the month-to-month changes for a specific month during a series of years and to judge concerning the degree of regularity of the seasonal movements for that month. *The significance of the final index of seasonal variation depends upon its typical character.* Observation of any column in Table 76 shows whether there is a tendency for the relatives to group about some value and how closely. The greater the concentration in a given column, the greater the significance of the index for that month. The influence of extreme variants is minimized by the use of the median or an average of central items.

IV. A simpler method. A simpler method has been presented in a recent article by Helen D. Falkner.[2] Some of the essentials of this method have been already suggested.

(1) The first step eliminates the influence of secular trend by expressing the original monthly data as percentages of the corresponding monthly trend values. For pig-iron production 1903–1914 these percentages are found in Table 78, page 354, column (6), and may be used by the student to illustrate the method developed by Miss Falkner.

(2) The second step is to form an array or frequency distribution of these percentages for each of the twelve months of the year (similar to the arrays in Table 76). The purpose of this procedure is to examine the monthly percentage fluctuations about the trend in order *to discover whether there exists a seasonal movement which can be generalized for the period,* and, if so, to determine its character and to measure it.

Observation of the tendency of the percentages in each monthly array to concentrate about a specific value and of the differences in these values from month to month reveals the nature of the seasonal movement.

(3) The third step is to determine *representative values* from the arrays,

[1] The student is urged to read "Correlation of Time Series," by Professor Persons, in the *Handbook of Mathematical Statistics*, chapter 10.

[2] "The Measurement of Seasonal Variation," Helen D. Falkner, *Journal of the American Statistical Association*, June, 1924.

for each of the twelve months. This is done by averaging *a number of central items in each array*, instead of all the items, thus avoiding the influence of extreme variants. The number of central items to be averaged is a *matter of experimentation* (ranging from three to seven), and depends upon the concentration of the items in the arrays. In any case extreme variants are excluded.

(4) Finally, the twelve typical monthly percentages which are the *crude indexes* of seasonal variation *are adjusted to average* 100 *per cent for a normal year* in the manner already illustrated in Table 74.

It is clear that this method has the advantages of simplicity and ease of computation. The experimental approach is valuable in the construction of a seasonal index. In doubtful cases various methods should be tried. Different time series do not permit the construction of equally reliable indexes of seasonal variation, by whatever method they may have been obtained.

Applying the method to pig-iron production 1903–1914. and using the percentages in Table 78, column (6), to form the monthly arrays, we obtain adjusted indexes of seasonal variation which differ somewhat more from those obtained by either Method II or III than those obtained by Methods II and III differ from each other.

TABLE 77. SUMMARY OF THE RESULTS OF THREE METHODS OF MEASURING SEASONAL VARIATION

MONTH	MONTHLY MEANS		TWELVE-MONTH MOVING AVERAGE (Table 74)	LINK RELATIVES (TABLE 76)	DIFFERENCE BETWEEN (4) AND (5)
	Trend neglected (Table 71)	Related to trend (Table 72)			
(1)	(2)	(3)	(4)	(5)	(6)
Jan......	95.8	95.6	99.2	100.0	.8
Feb.....	92.8	92.3	94.1	93.7	.4
Mar.....	105.8	105.0	106.0	106.1	.1
April....	103.5	102.4	103.4	103.1	.3
May	104.2	102.8	104.2	103.8	.4
June	98.5	96.9	98.3	97.1	1.2
July.....	97.7	99.4	97.0	95.8	1.2
Aug.....	100.3	101.7	98.5	98.0	.5
Sept.....	101.1	102.2	98.1	97.7	.4
Oct......	106.2	107.0	103.8	104.2	.4
Nov.....	98.3	98.7	98.2	99.4	1.2
Dec.....	96.0	96.2	99.1	100.8	1.7
Total....	1200.2	1200.2	1199.9	1199.7	
Mean....	100.0	100.0	100.0	100.0	

Observation of Table 77 and Figure 54 seems to show a closer correspondence in results obtained from the method based on a twelve-month

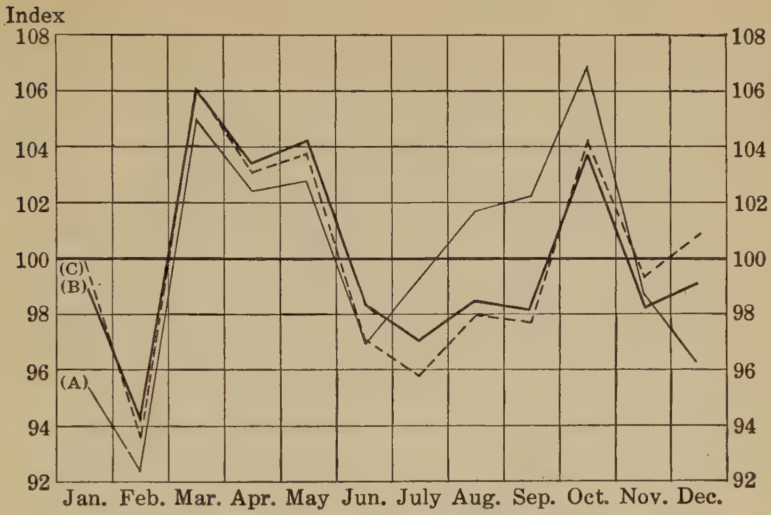


FIG. 54. COMPARISON OF THREE INDEXES OF SEASONAL VARIATION CONSTRUCTED BY DIFFERENT METHODS

(Data from Table 77, columns (3), (4) and (5).)

Key: (A) = Monthly Means Related to Trend, column (3), Table 77.
(B) = Method Based on Twelve-Month Moving Averages, column (4), Table 77.
(C) = Method of Link Relatives, column (5), Table 77.

moving average and the method of link relatives, columns (4) and (5), than in the results obtained from either of these methods compared with the results of the method of monthly means, especially during the latter half of the year.

It is logical to conclude that for certain types of data and for specific periods of time the method of monthly means, which is the simplest to apply, will yield results corresponding more closely with other methods than those shown in the case of pig-iron production 1903–1914. This would be true where there exists little or no secular trend and where the cycles are moderate and regularly distributed over the period, with no extreme variations due to non-seasonal influences. *Where there is a lack of homogeneity in the seasonal movements of the various years of the period any index has less significance.* Close scrutiny of the columns of Tables 74 and 76 illustrate this statement and suggest why the indexes shown in Figure 54 differ more for certain months than for others. Where extreme variations exist the method of monthly means is less reliable than the other methods which use the median, or an average of middle values in the series for obtaining the monthly types. *It is desirable to experiment with different methods applied to a given kind of data and to a specific period.*

Table 78. Elimination of Seasonal Variations — Cycles 1903–1914

Monthly Pig-Iron Production — Twelve-Month Moving Averages Centered — Monthly Trend Values — Index of Seasonal Variation Applied — Cycles in Percentage Deviation from the Trend — Cycles in Units of Standard Deviation (σ)

Year and Month	Original data (unit = 1000 long tons)	12-month moving average centered (Table 73)	Percentages (2) ÷ (3) × 100	Monthly trend values (monthly increment = 6 units)	Percentage (original to trend) (2) ÷ (5) × 100	Index of seasonal variation (Table 77)	Column (7) subtracted from column (6) (Cycles = x)	Cycles in percentages squared x^2	Cycles in units of σ. (8) ÷ 18.9[a] $\frac{x}{\sigma}$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1903 Jan.	1472	1519	96.9	1519	96.9	99.2	− 2.3	5.29	− .1
Feb.	1390	1526	91.1	1525	91.1	94.1	− 3.0	9.00	− .2
Mar.	1590	1537	103.4	1531	103.9	106.0	− 2.1	4.41	− .1
Apr.	1608	1540	104.4	1537	104.6	103.4	1.2	1.44	.1
May	1713	1521	112.6	1543	111.0	104.2	6.8	46.24	.4
June	1673	1478	113.2	1549	108.0	98.3	9.7	94.09	.5
July	1546	1429	108.2	1555	99.4	97.0	2.4	5.76	.1
Aug.	1571	1399	112.3	1561	100.6	98.5	2.1	4.41	.1
Sept.	1553	1385	112.1	1567	99.1	98.1	1.0	1.00	.1
Oct.	1425	1377	103.5	1573	90.6	103.8	−13.2	174.24	− .7
Nov.	1039	1367	76.0	1579	65.8	98.2	−32.4	1049.76	−1.7
Dec.	846	1344	62.9	1585	53.4	99.1	−45.7	2088.49	−2.4
1904 Jan.	1921	1310	70.3	1591	57.9	99.2	−41.3	1705.69	−2.2
Feb.	1205	1274	94.6	1597	75.5	94.1	−18.6	345.96	−1.0
Mar.	1447	1250	115.8	1603	90.3	106.0	−15.7	246.49	− .8
Apr.	1555	1242	125.2	1609	96.6	103.4	− 6.8	46.24	− .4
May	1534	1261	121.6	1615	95.0	104.2	− 9.2	84.64	− .5
June	1292	1312	98.5	1621	79.7	98.3	−18.6	345.96	−1.0
July	1106	1380	80.1	1627	68.0	97.0	−29.0	841.00	−1.5
Aug.	1167	1433	81.4	1633	71.5	98.5	−27.0	729.00	−1.4
Sept.	1352	1469	92.0	1639	82.5	98.1	−15.6	243.36	− .8
Oct.	1450	1504	96.4	1645	88.1	103.8	−15.7	246.49	− .8
Nov.	1486	1538	96.6	1651	90.0	98.2	− 8.2	67.24	− .4
Dec.	1616	1577	102.5	1657	97.5	99.1	− 1.6	2.56	− .1

[a] The value of σ is 18.9 per cent computed from column (9). Table continued on following pages.

TABLE 78. ELIMINATION OF SEASONAL VARIATIONS (*continued*)

Year and Month (1)	Original data (unit = 1000 long tons) (2)	12-month moving average centered (Table 73) (3)	Percentages (2) ÷ (3) × 100 (4)	Monthly trend values (monthly increment = 6 units) (5)	Percentage (original to trend) (2) ÷ (5) × 100 (6)	Index of seasonal variation (Table 77) (7)	Column (7) subtracted from column (6). (Cycles = x) (8)	Cycles in percentages squared x^2 (9)	Cycles in units of σ. (8) ÷ 18.9 $\frac{x}{\sigma}$ (10)
1905 Jan.	1781	1623	109.7	1663	107.1	99.2	7.9	62.41	.4
Feb.	1597	1679	95.1	1669	95.7	94.1	1.6	2.56	.1
Mar.	1936	1729	112.0	1675	115.6	106.0	9.6	92.16	.5
Apr.	1922	1778	108.1	1681	114.3	103.4	10.9	118.81	.6
May	1963	1825	107.6	1687	116.4	104.2	12.2	148.84	.6
June	1793	1864	96.2	1693	105.9	98.3	7.6	57.76	.4
July	1741	1894	91.9	1699	102.5	97.0	5.5	30.25	.3
Aug.	1843	1919	96.0	1705	108.1	98.5	9.6	92.16	.5
Sept.	1899	1941	97.8	1711	111.0	98.1	12.9	166.41	.7
Oct.	2053	1957	104.9	1717	119.6	103.8	15.8	249.64	.8
Nov.	2014	1968	102.3	1723	116.9	98.2	18.7	349.69	1.0
Dec.	2045	1982	103.2	1729	118.3	99.1	19.2	368.64	1.0
1906 Jan.	2068	2000	103.4	1735	119.2	99.2	20.0	400.00	1.1
Feb.	1904	2016	94.4	1741	109.4	94.1	15.3	234.09	.8
Mar.	2155	2021	106.6	1747	123.4	106.0	17.4	302.76	.9
Apr.	2073	2030	102.1	1753	118.3	103.4	14.9	222.01	.8
May	2098	2043	102.7	1759	119.3	104.2	15.1	228.01	.8
June	1976	2058	96.0	1765	112.0	98.3	13.7	187.69	.7
July	2013	2072	97.2	1771	113.7	97.0	16.7	278 89	.9
Aug.	1926	2083	92.5	1777	108.4	98.5	9.9	98.01	.5
Sept.	1960	2092	93.7	1783	109.9	98.1	11.8	139.24	.6
Oct.	2196	2101	104.5	1789	122.8	103.8	19.0	361.00	1.0
Nov.	2187	2115	103.4	1795	121.8	98.2	23.6	556.96	1.2
Dec.	2235	2134	104.7	1801	124.1	99.1	25.0	625.00	1.3

TABLE 78. ELIMINATION OF SEASONAL VARIATIONS (*continued*)

Year and Month (1)	Original data (unit = 1000 long tons) (2)	12-month moving average centered (Table 73) (3)	Percentages (2) ÷ (3) × 100 (4)	Monthly trend values (monthly increment = 6 units) (5)	Percentage (original to trend) (2) ÷ (5) × 100 (6)	Index of seasonal variation (Table 77) (7)	Column (7) subtracted from column (6). (Cycles = x) (8)	Cycles in percentages squared x^2 (9)	Cycles in units of σ. (8) ÷ 18.9 $\frac{x}{\sigma}$ (10)
1907 Jan.	2205	2155	102.3	1807	122.0	99.2	22.8	519.84	1.2
Feb.	2045	2178	93.9	1813	112.8	94.1	18.7	349.69	1.0
Mar.	2226	2202	101.1	1819	122.4	106.0	16.4	268.96	.9
Apr.	2216	2216	100.0	1825	121.4	103.4	18.0	324.00	1.0
May	2295	2207	104.0	1831	125.3	104.2	21.1	445.21	1.1
June	2234	2151	103.9	1837	121.6	98.3	23.3	542.89	1.2
July	2255	2060	109.5	1843	122.4	97.0	25.4	645.16	1.3
Aug.	2250	1972	114.1	1849	121.7	98.5	23.2	538.24	1.2
Sept.	2183	1890	115.5	1855	117.7	98.1	19.6	384.16	1.0
Oct.	2336	1804	129.5	1861	125.5	103.8	21.7	470.89	1.1
Nov.	1828	1713	106.7	1867	97.9	98.2	− .3	.09	− .0[a]
Dec.	1234	1617	76.3	1873	65.9	99.1	−33.2	1102.24	−1.8
1908 Jan.	1045	1527	68.4	1879	55.6	99.2	−43.6	1900.96	−2.3
Feb.	1077	1447	74.4	1885	57.1	94.1	−37.0	1369.00	−2.0
Mar.	1228	1377	89.2	1891	64.9	106.0	−41.1	1689.21	−2.2
Apr.	1149	1312	87.6	1897	60.6	103.4	−42.8	1831.84	−2.3
May	1165	1270	91.7	1903	61.2	104.2	−43.0	1849.00	−2.3
June	1092	1281	85.2	1909	57.2	98.3	−41.1	1689.21	−2.2
July	1218	1334	91.3	1915	63.6	97.0	−33.4	1115.56	−1.8
Aug.	1348	1391	96.9	1921	70.2	98.5	−28.3	800.89	−1.5
Sept.	1418	1442	98.3	1927	73.6	98.1	−24.5	600.25	−1.3
Oct.	1563	1491	104.8	1933	80.9	103.8	−22.9	524.41	−1.2
Nov.	1577	1546	102.0	1939	81.3	98.2	−16.9	285.61	− .9
Dec.	1740	1611	108.0	1945	89.5	99.1	− 9.6	92.16	− .5
Average for 12 years =				1948					

[a] Less than one tenth σ.

TABLE 78. ELIMINATION OF SEASONAL VARIATIONS (*continued*)

Year and Month		Original data (unit = 1000 long tons)	12-month moving average centered (Table 73)	Percentages (2) ÷ (3) × 100	Monthly trend values (monthly increment = 6 units)	Percentage (original to trend) (2) ÷ (5) × 100	Index of seasonal variation (Table 77)	Column (7) subtracted from column (6) (Cycles = x)	Cycles in percentages squared x^2	Cycles in units of σ. (8) ÷ 18.9 $\frac{x}{\sigma}$
(1)		(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1909	Jan.	1801	1683	107.0	1951	92.3	99.2	− 6.9	47.61	− .4
	Feb.	1703	1756	97.0	1957	87.0	94.1	− 7.1	50.41	− .4
	Mar.	1832	1835	99.8	1963	93.3	106.0	−12.7	161.29	− .7
	Apr.	1738	1918	90.6	1969	88.3	103.4	−15.1	228.01	− .8
	May	1880	2001	94.0	1975	95.2	104.2	− 9.0	81.00	− .5
	June	1929	2079	92.8	1981	97.4	98.3	− .9	.81	− .0[a]
	July	2101	2150	97.7	1987	105.7	97.0	8.7	75.69	.5
	Aug.	2246	2213	101.5	1993	112.7	98.5	14.2	201.64	.8
	Sept.	2385	2275	104.8	1999	119.3	98.1	21.2	449.44	1.1
	Oct.	2600	2338	111.2	2005	129.7	103.8	25.9	670.81	1.4
	Nov.	2547	2390	106.6	2011	126.7	98.2	28.5	812.25	1.5
	Dec.	2635	2426	108.6	2017	130.6	99.1	31.5	992.25	1.7
1910	Jan.	2608	2442	106.8	2023	128.9	99.2	29.7	882.09	1.6
	Feb.	2397	2437	98.4	2029	118.1	94.1	24.0	576.00	1.3
	Mar.	2617	2418	108.2	2035	128.6	106.0	22.6	510.76	1.2
	Apr.	2483	2383	104.2	2041	121.7	103.4	18.3	334.89	1.0
	May	2390	2336	102.3	2047	116.8	104.2	12.6	158.76	.7
	June	2265	2273	99.6	2053	110.3	98.3	12.0	144.00	.6
	July	2148	2202	97.5	2059	104.3	97.0	7.3	53.29	.4
	Aug.	2106	2141	98.4	2065	102.0	98.5	3.5	12.25	.2
	Sept.	2056	2099	98.0	2071	99.3	98.1	1.2	1.44	.1
	Oct.	2093	2063	101.5	2077	100.8	103.8	− 3.0	9.00	− .2
	Nov.	1909	2025	94.3	2083	91.6	98.2	− 6.6	43.56	− .3
	Dec.	1777	1985	89.5	2089	85.1	99.1	−14.0	196.00	− .7

[a] Less than one tenth σ.

TABLE 78. ELIMINATION OF SEASONAL VARIATIONS (*continued*)

Year and Month (1)	Original data (unit = 1000 long tons) (2)	12-month moving average centered (Table 73) (3)	Percentages (2) ÷ (3) × 100 (4)	Monthly trend values (monthly increment = 6 units) (5)	Percentage (original to trend) (2) ÷ (5) × 100 (6)	Index of seasonal variation (Table 77) (7)	Column (7) subtracted from column (6) (Cycles = x) (8)	Cycles in percentages squared x^2 (9)	Cycles in units of σ. (8) ÷ 18.9 $\frac{x}{\sigma}$ (10)
1911 Jan.	1759	1950	90.2	2095	84.0	99.2	−15.2	231.04	− .8
Feb.	1794	1927	93.1	2101	85.4	94.1	− 8.7	75.69	− .5
Mar.	2188	1917	114.1	2107	103.8	106.0	− 2.2	4.84	− .1
Apr.	2065	1913	107.9	2113	97.7	103.4	− 5.7	32.49	− .3
May	1893	1918	98.7	2119	89.3	104.2	−14.9	222.01	− .8
June	1787	1933	92.4	2125	84.1	98.3	−14.2	201.64	− .8
July	1793	1957	91.6	2131	84.1	97.0	−12.9	166.41	− .7
Aug.	1926	1981	97.2	2137	90.1	98.5	− 8.4	70.56	− .4
Sept.	1977	2003	98.7	2143	92.3	98.1	− 5.8	33.64	− .3
Oct.	2102	2025	103.8	2149	97.8	103.8	− 6.0	36.00	− .3
Nov.	1999	2064	96.9	2155	92.8	98.2	− 5.4	29.16	− .3
Dec.	2043	2117	96.5	2161	94.5	99.1	− 4.6	21.16	− .2
1912 Jan.	2057	2170	94.8	2167	94.9	99.2	− 4.3	18.49	− .2
Feb.	2100	2220	94.6	2173	96.6	94.1	2.5	6.25	.1
Mar.	2405	2265	106.2	2179	110.4	106.0	4.4	19.36	.2
Apr.	2375	2309	102.9	2185	108.7	103.4	5.3	28.09	.3
May	2512	2360	106.4	2191	114.7	104.2	10.5	110.25	.6
June	2440	2417	101.0	2197	111.1	98.3	12.8	163.84	.7
July	2410	2479	97.2	2203	109.4	97.0	12.4	153.76	.7
Aug.	2512	2529	99.3	2209	113.7	98.5	15.2	231.04	.8
Sept.	2463	2565	96.0	2215	111.2	98.1	13.1	171.61	.7
Oct.	2689	2596	103.6	2221	121.1	103.8	17.3	299.29	.9
Nov.	2630	2624	100.2	2227	118.1	98.2	19.9	396.01	1.1
Dec.	2782	2645	105.2	2233	124.6	99.1	25.5	650.25	1.3

TABLE 78. ELIMINATION OF SEASONAL VARIATIONS (*continued*)

Year and Month	Original data (unit = 1000 long tons)	12-month moving average centered (Table 73)	Percentages (2) ÷ (3) × 100	Monthly trend values (monthly increment = 6 units)	Percentage (original to trend) (2) ÷ (5) × 100	Index of seasonal variation (Table 77)	Column (7) subtracted from column (6) (Cycles = x)	Cycles in percentages squared x^2	Cycles in units of σ. (8) ÷ 18.9 $\frac{x}{\sigma}$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1913 Jan.	2795	2659	105.1	2239	124.8	99.2	25.6	655.36	1.4
Feb.	2586	2666	97.0	2245	115.2	94.1	21.1	445.21	1.1
Mar.	2763	2670	103.5	2251	122.7	106.0	16.7	278.89	.9
Apr.	2752	2665	103.3	2257	121.9	103.4	18.5	342.25	1.0
May	2822	2642	106.8	2263	124.7	104.2	20.5	420.25	1.1
June	2628	2593	101.3	2269	115.8	98.3	17.5	306.25	.9
July	2560	2522	101.5	2275	112.5	97.0	15.5	240.25	.8
Aug.	2543	2455	103.6	2281	111.5	98.5	13.0	169.00	.7
Sept.	2505	2409	104.0	2287	109.5	98.1	11.4	129.96	.6
Oct.	2546	2371	107.4	2293	111.0	103.8	7.2	51.84	.4
Nov.	2233	2320	96.3	2299	97.1	98.2	− 1.1	1.21	− .1
Dec.	1983	2261	87.7	2305	86.0	99.1	−13.1	171.61	− .7
1914 Jan.	1885	2206	85.4	2311	81.6	99.2	−17.6	309.76	− .9
Feb.	1888	2158	87.5	2317	81.5	94.1	−12.6	158.76	− .7
Mar.	2348	2109	111.3	2323	101.1	106.0	− 4.9	24.01	− .3
Apr.	2270	2051	110.7	2329	97.5	103.4	− 5.9	34.81	− .3
May	2093	1989	105.2	2335	89.6	104.2	−14.6	213.16	− .8
June	1918	1941	98.8	2341	81.9	98.3	−16.4	268.96	− .9
July	1958	1909	102.6	2347	83.4	97.0	−13.6	184.96	− .7
Aug.	1995	1888	105.7	2353	84.8	98.5	−13.7	187.69	− .7
Sept.	1883	1867	100.9	2359	79.8	98.1	−18.3	334.89	−1.0
Oct.	1778	1850	96.1	2365	75.2	103.8	−28.6	817.96	−1.5
Nov.	1518	1850	82.1	2371	64.0	98.2	−34.2	1169.64	−1.8
Dec.	1516	1876	80.8	2377	63.8	99.1	−35.3	1246.09	−1.9

Summation of Column (9) for the entire Table 78 = 51,330.12

$$\sigma^2 = \frac{\Sigma x^2}{N}$$

$$\sigma^2 = \frac{51{,}330.12}{144}$$

$$\sigma = \sqrt{\frac{51{,}330.12}{144}}$$

σ = 18.9 per cent

In the case of pig-iron production 1903–1914, the choice of methods seems to lie between that based upon a twelve-month moving average and the method of link relatives. Since the results obtained by the two methods do not differ greatly for most of the months, *either index may be used to eliminate the seasonal variation from the original data.* The method of construction based upon the twelve-month moving average is somewhat more easily understood. *No matter how obtained, the method of application of the various indexes is the same.* To illustrate the method of applying the index of seasonal variation the monthly indexes from Table 77, column (4), are used for each year during the entire period 1903–1914 (Method II, Table 74).

APPLICATION OF THE INDEX OF SEASONAL VARIATION

Methods for measuring secular trend and seasonal movement have been described and illustrated. It remains to show how the influence of these factors may be eliminated from the original figures. Table 78 shows the procedure and presents the final monthly values which *measure the cyclical movements, practically undisturbed by the other factors.*

In column (2) of Table 78 are given the monthly production figures and in column (5) the monthly values for the secular trend. The trend was determined, on the basis of the annual average monthly production, by fitting a straight line by the method of least squares, as fully explained in the early part of the chapter. The annual growth in production was shown in Table 65, page 323, to be 72 unit-tons between 1903 and 1914. The monthly increment is $\frac{1}{12}$ of 72 unit-tons, which equals 6 unit-tons. The average monthly production for the twelve years was 1948 unit-tons, which was plotted between December 1908 and January 1909, the middle of the period (Figure 44). Taking this point as the origin, it is necessary to subtract one half of the monthly increment, 3 unit-tons, from 1948 to obtain a figure for December 1908 which is comparable with the original item for that month in column (2), located at the middle of the month. Likewise, 3 unit-tons must be added to 1948 to obtain a value for mid-January 1909. Successive monthly values for the ordinates of secular trend are obtained by subtracting or adding 6 unit-tons until the 144 monthly values are established, located on the straight line in Figure 44. This procedure merely locates twelve values for each year in place of the single annual value representing the trend. (See page 326.)

In column (6) of the table the ratio of each of the original monthly items to the corresponding trend value is expressed as a percentage. In other words, the influence of the secular trend upon the cycles is eliminated by regarding it as 100 in each case, and by measuring the deviations above and below it in terms of percentage. This is exactly the same procedure as illustrated in Figure 48 for the annual data. Since the secular trend is the base from which the cyclical deviations are measured it may be represented as a horizontal straight line, about which the monthly fluctuations are shown in Figure 55, page 362. *The secular trend or growth factor has been eliminated by this procedure.*

Elimination of seasonal fluctuations. In the percentages plotted in Figure 55 the influence of the seasonal movement is still combined with the cyclical. *It remains to eliminate this movement from the cycles.* The index describing this type of variation, Table 77,[1] column (4), is entered in column (7) of Table 78, *repeated for each of the* 12 *years*, because this index has been constructed from the experience of the entire period and *represents not a particular year but a normal year.*

It will assist the student to visualize the influence of the seasonal movement upon the original monthly data related to the secular trend, if we plot the seasonal indexes as deviations above and below the horizontal line, which represents the secular trend. In Figure 56, therefore, the seasonal indexes are shown, plotted as deviations about the trend line regarded as 100, and in the same diagram the original data are related to the trend values in terms of percentages (Columns (6) and (7) in Table 78).

The next step is actually to eliminate the seasonal movement by subtracting the percentages of column (7) from those of column (6), *carefully preserving the proper signs of the remainders. The resulting series of positive and negative percentage deviations from the trend, in column* (8), *describes the cyclical movements, undisturbed by secular trend or by seasonal variations.* These cycles are represented in Figure 57, page 363.

It should be apparent from Figure 56, that, when the seasonal swing is in the same direction as the cyclical movement (on the same side of the trend line) for a given month, the cyclical deviation from the trend is reduced by subtracting the percentage in column (7) Table 78 from the corresponding percentage in (6) which represents the deviation of the original data from the trend, *a part of which is due to the seasonal movement in the same direction.* On the contrary, when the seasonal swing is in the opposite direction from the cyclical movement (on opposite sides of the trend line), subtracting the seasonal index for the given month, column (7), from the percentage deviation of the original data from the trend, column (6), really increases the cyclical movement from

[1] The method of applying the seasonal index is the same for all the indexes given in Table 77.

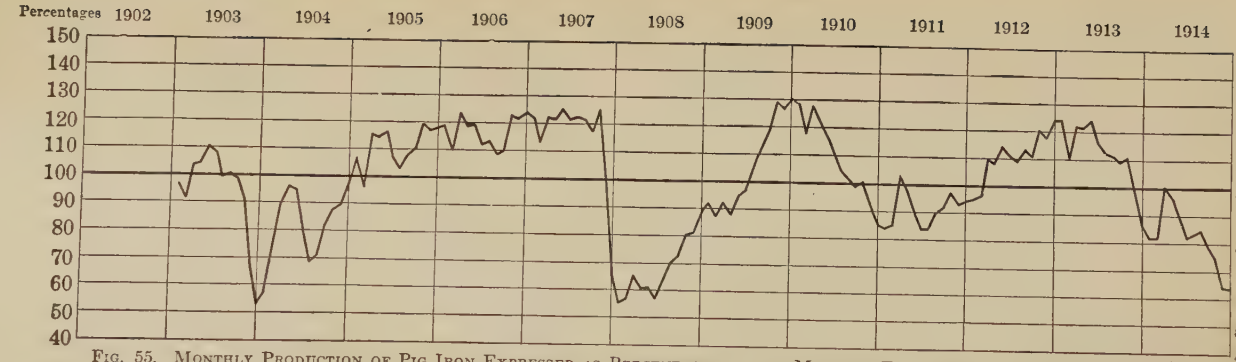


FIG. 55. MONTHLY PRODUCTION OF PIG IRON EXPRESSED AS PERCENTAGES OF THE MONTHLY TREND VALUES (TREND VALUES = 100) Cyclical fluctuations are shown with seasonal factor included in the data. Secular trend eliminated. (Data from Table 78, column (6).)

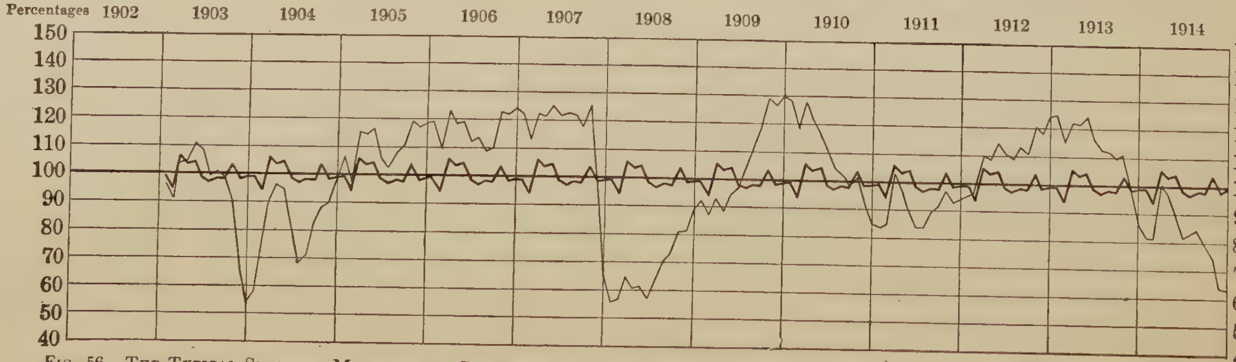


FIG. 56. THE TYPICAL SEASONAL MOVEMENT AS REPRESENTED BY AN INDEX OF SEASONAL VARIATION PLOTTED FOR EACH MONTH ABOVE AND BELOW THE TREND LINE AS 100

The relations of the original monthly items to the monthly trend values are shown as in Fig. 55. (Data for seasonal indexes from Table 78, column (7). The lighter line is identical with Fig. 55.)

the trend, *which has been retarded by the seasonal movement in the opposite direction.* These statements may be verified by comparing the distances between the two curves in Figure 56 with the values in columns (6) and (7) in Table 78 for specific months.

CYCLES EXPRESSED IN UNITS OF σ

Column (8) of Table 78 describes the cycles in percentage deviations from the trend. Column (9) records the squares of the values in column (8). It is only necessary to add the values of column (9), divide by the number of monthly items (144), and extract the square root of this quotient to find the standard deviation of the entire series of percentage deviations, 18.9 per cent. We wish to express the percentage deviations of column (8) each in terms of the unit σ. Therefore, each monthly item of column (8) is divided by 18.9 and the quotients are entered in column (10).

Setting forth the cycles in the form of percentages or units of σ renders the data independent of the particular units in which the original values are expressed, and permits us to combine more than one series into a composite curve describing cyclical variations, as illustrated in Figure 50.

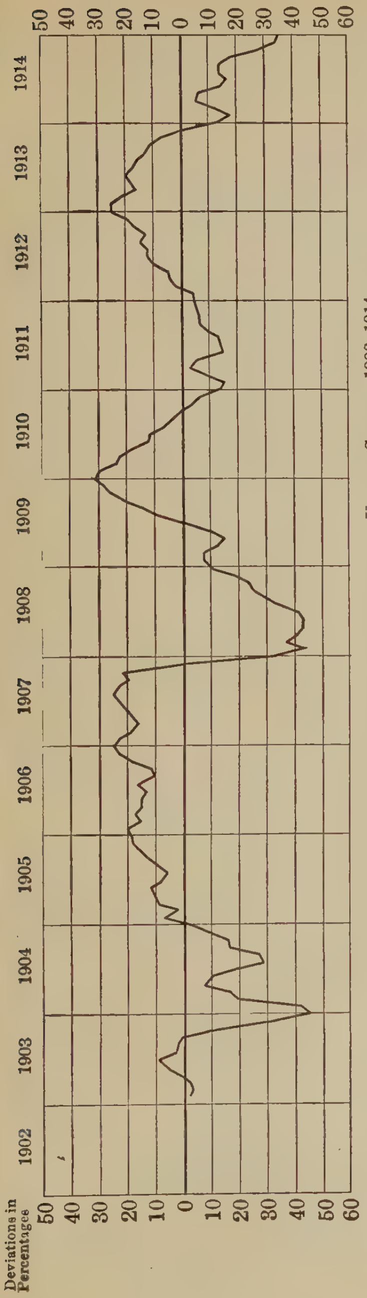


Fig. 57. Cycles of Pig-Iron Production in the United States, 1903–1914

The seasonal factor and the secular trend have been eliminated. The irregular and accidental factors are not eliminated. (Data from Table 78, column (8). Column (10) of Table 78 expresses the deviations from the trend in units of the standard deviation, σ. These may be plotted for each month, as in Fig. 49.)

The common units of σ also make possible comparison of pig-iron production with other series treated in a similar manner, in respect to the relative shapes of the curves and the presence and amount of lag.

IMPORTANCE OF GRAPHIC COMPARISON

Comparison of the graphic representations of time series is essential to a complete understanding of them. Summary figures do not describe the data adequately. They describe only average conditions. Relationships between two or more time series are more completely described by the use of graphic devices, in addition to such measures as correlation coefficients. The cycles of two or more series, each expressed in units of their respective σ's, may be compared by drawing the curves on separate sheets and by placing one curve over another against a window pane or on a ground-glass surface beneath which is placed an electric light. Differences are easily noted and it is possible to detect the presence of lag and to estimate the approximate period by which movements in one series fall behind or lead those of another.

CORRELATION OF SERIES EXPRESSED IN UNITS OF σ

It is very simple to correlate directly the σ units of deviation in column (10) with similar units of another series. It is necessary merely to take the algebraic summation of the xy products of the two series in units of σ, as expressed in column (10), Table 78, and to divide by the number of pairs correlated, in this case 144 pairs,

$$\frac{\Sigma\left(\frac{x}{\sigma}\text{ times }\frac{y}{\sigma}\right)}{N}$$

The xy deviations are already in terms of their respective standard deviations. This procedure is illustrated for the annual data in Table 68, column (11), page 331. No new method is required for the monthly items.

MEASURING THE LAG FROM MONTHLY DATA

Having discovered the presence of lag from a comparison of the curves of two time series, we can measure the amount by correlating different pairs of the two series. Monthly data make possible a more exact measurement than could be made in the early part of the chapter where only annual figures were used. However, the method is exactly the same as that shown in Table 70, page 338.

For example, monthly interest rates may be analyzed in the same manner as described in the preceding pages for pig-iron production. Having secured the cycles for both series, corrected for secular trends and seasonal variation, as in column (8) or column (10) of Table 78, we can correlate February interest-rate deviations with January pig-iron deviations; March interest-rate deviations with January pig-iron; April interest-rate deviations with January pig-iron, etc. A coefficient of correlation may be obtained for each arrangement of pairs. This procedure amounts to moving the interest-rate series backward one, two, three, or more months, on the assumption of different periods of lag, relating the series each time to pig-iron production. *The object is to secure the closest correspondence between the movements in the two series — in graphic terms, the best fit of the two curves to each other.*

If and as long as the coefficient of correlation increases in size from the different pairings of monthly items a progressive improvement in the correspondence in the movements of the two series is indicated. When the coefficient reaches a maximum and begins to decrease in size from successive pairings then it may be concluded that the period of lag is measured by the period of time over which the lagging series has been moved in order to obtain the maximum degree of correspondence in the movements of the two series.

For example, if the July interest-rate deviations, when paired with the preceding January pig-iron deviations, a lag of six months throughout the entire series, produce the highest coefficient of correlation the conclusion is that the movement in interest rates takes place about six months after the corresponding movement in pig-iron production. The method of making these correlations and measuring the lag is illustrated in Tables 68 and 70, where annual data were used. No new principle is involved in the use of the monthly data, after they have been corrected for seasonal and secular changes, as in columns (8) or (10) of Table 78. The only difference is that there are 144 items in each series instead of 12.

Utility of the measurements of lag in forecasting. If the relationships described in this chapter can be shown to exist between two or more time series, and if the approximate lag can be measured over a period of sufficient length to establish confidence in the reality of the recurring movements, it becomes possible to use our past and present knowledge of one or more series to forecast probable changes in the related series at approximate dates *in the near future.*

A problem proposed. In Table 79 is found the average monthly interest rates on 60 to 90 day commercial paper in New York City, 1902–1915. The student is urged to use this time series for practice in constructing a seasonal index, and to relate the cycles of pig-iron production and interest rates, after each series has been corrected so as to eliminate the other disturbing factors. The data should be tested for lag and this should be measured in the manner suggested.

TABLE 79. AVERAGE MONTHLY RATES OF INTEREST ON 60–90 DAY COMMERCIAL PAPER IN NEW YORK, 1902–1915 [a]

(Unit one per cent)

MONTH	1902	1903	1904	1905	1906	1907	1908	1909	1910	1911	1912	1913	1914	1915
Jan........	4.56	5.22	4.89	4.00	5.06	6.15	6.59	3.68	4.75	3.98	3.90	4.93	4.53	3.84
Feb.......	4.00	4.90	4.79	3.81	5.04	5.94	5.06	3.54	4.44	4.09	3.75	4.91	3.84	3.75
Mar.......	4.37	5.54	4.68	3.93	5.28	6.19	5.63	3.50	4.50	3.88	4.19	5.75	3.88	3.38
April	4.53	5.19	4.13	4.00	5.44	5.92	4.38	3.50	4.75	3.66	4.15	5.53	3.73	3.66
May......	4.54	4.75	3.93	3.98	5.33	5.40	3.94	3.44	4.75	3.63	4.19	5.36	3.88	3 72
June......	4.42	5.16	3.60	3.75	5.25	5.50	3.69	3.25	4.81	3.69	4.00	5.88	3.84	3.65
July	4.64	5.43	3.55	4.13	5.48	5.75	3.75	3.38	5.38	3.78	4.53	6.06	4.40	3.25
Aug	4.82	5.94	3.84	4.19	6.00	6.25	3.61	4.04	5.43	4.19	5.00	6.00	6.34	3.53
Sept......	5.58	6.00	4.29	4.72	6.56	6.79	3.89	4.25	5.53	4.54	5.56	5.78	6.70	3.25
Oct.......	5.90	5.79	4.41	4.92	6.30	7.10	4.10	5.03	5.56	4.35	5.93	5.69	6.44	3.22
Nov	5.71	5.95	4.14	5.53	6.25	7.40	4.04	5.09	5.50	3.91	5.72	5.56	5.50	2.98
Dec.......	6.00	5.79	4.28	5.79	6.25	8.00	3.85	5.09	4.66	4.63	6.00	5.68	4.35	3.13
Annual average	4.92	5.47	4.21	4.40	*5.68	*6.36	4.38	3.98	*5.00	4.03	4.74	*5.60	*4.78	3.45

[a] Data from *The Review of Economic Statistics*, Preliminary Volume I, p. 99, Committee on Economic Research, Harvard University. Revised data are given in *The Review of Economic Statistics*, January, 1923, pp. 28–29.

* If the student computes the annual averages in the last row of the table by hand or by machine he will find a slight discrepancy in the second decimal place in the figures marked (*). This is due to the fact that the Harvard Committee staff used the slide rule to compute the averages. The author desires to use the figures as given in the above reference in order not to confuse the student in the use of this very important source of data on the time series. The annual averages found in Wesley C. Mitchell's *Business Cycles*, pp. 166–67, differ somewhat from these.

Summary. To describe and to measure the relationships between time series, where monthly or quarterly data are available, the following procedures are necessary.

(1) Testing each series for secular trend and determining the trend, if present.

(2) Testing each series for the presence of seasonal variation, and constructing an index to measure this movement, if present.

(3) Correcting the original items for seasonal and secular variations.

(4) Expressing the corrected deviations in graphic form, usually in terms of percentage deviations or in units of σ, in order to compare the movements of the two or more series in detail, and to estimate the probable amount of lag, if present.

(5) Computing correlation coefficients on the assumption of different periods of lag, in order to measure the amount of the lag or lead more exactly.

READINGS

Persons, W. M., "Correlation of Time Series," *Handbook of Mathematical Statistics* (H. L. Rietz, Editor), chap. 10. (See also *Journal of American Statistical Association*, June, 1923.)

Mills, F. C., *Statistical Methods As Applied to Economics and Business*, chap. 7 ("Trend"), chap. 8 ("Seasonal and Cyclical Fluctuations"), chap. 11 ("Relationship between Time Series").

King, W. I., *Elements of Statistical Method*, chap. 15.

Jerome, Harry, *Statistical Method*, chaps. 13 and 14.

Davies, G. R., *Introduction to Economic Statistics*, chap. 5.

Statistical Analysis and Projection of Time Series, Statistical Bulletin No. 4, Statistical Method Series, American Telephone and Telegraph Company, Office of the Chief Statistician, 1922. (Excellent application of methods.)

Yule, G. U., *An Introduction to the Theory of Statistics*, 6th ed., chap. 10.

Hart, W. L., "The Method of Monthly Means for Determination of a Seasonal Variation," *Journal of the American Statistical Association*, September, 1922.

Falkner, Helen D., "The Measurement of Seasonal Variation," *Journal of the American Statistical Association*, June, 1924.

Hall, L. W., "Seasonal Variations as a Relative of Secular Trend," *Journal of the American Statistical Association*, June, 1924.

King, W. I., "An Improved Method for Measuring the Seasonal Factor," *Journal of the American Statistical Association*, September, 1924. (By this method changes in the seasonal factor from year to year are taken into consideration.)

Persons, W. M., *The Review of Economic Statistics*, Committee on Economic Statistics, Harvard University, Preliminary Volume I, 1919. (An explanation of methods useful in the analysis of time series and their application to many series. An excellent source for practice problems.)

REFERENCES

Moore, Henry L., *Economic Cycles: Their Law and Cause.*

—— ——, *Forecasting the Yield and the Price of Cotton.*

—— ——, *Generating Economic Cycles.*

Mitchell, Wesley C., *Business Cycles.*

Jordan, D. F., *Business Forecasting.*

The Problem of Business Forecasting (edited by Persons, Foster and Hettinger), Publication of the Pollak Foundation for Economic Research. Papers presented at the Eighty-fifth Annual Meeting of the American Statistical Association, December, 1923, published 1924.

Persons, W. M., "Construction of a Business Barometer," *American Economic Review*, December, 1916. (Based upon annual data.)

—— ——, "The Variate Difference Correlation Method and Curve Fitting," *Quarterly Publication of the American Statistical Association*, June, 1917.

Yule, G. U., "On the Time Correlation Problem, with Especial Reference to the Variate Difference Correlation Method," *Journal of the Royal Statistical Society*, July, 1921.

Pearl, Raymond, *Medical Biometry and Statistics*, chap. 16 ("Simple Curve Fitting").

Karsten, Karl G., *Charts and Graphs*, chap. 42 ("Curve Fitting").

Huntington, E. V., "Curve Fitting by the Method of Least Squares and the Method of Moments," *Handbook of Mathematical Statistics* (H. L. Rietz, Editor), chap. 4.

Merriman, M., *A Textbook on the Method of Least Squares.*

Weld, L. D., *Theory of Errors and Least Squares.*

[Details as to publisher and date of publication of the readings and references given at the close of each chapter are found in the alphabetical list in Appendix A.]

PART III

THE GATHERING AND PRESENTATION OF STATISTICAL DATA

When the Statistical Society of London, now the Royal Statistical Society, was organized in 1834, its functions, as stated in the prospectus, were " to procure, arrange, and publish facts calculated to illustrate the condition and prospects of Society."

(*The History of Statistics*, published for the American Statistical Association by The Macmillan Company, 1918, p. 385.)

CHAPTER XIV

COLLECTION OF STATISTICAL DATA

A PRELIMINARY STUDY OF THE PROBLEM

The first step in the planning of an original statistical investigation is to bring together existing information, in as much detail as possible, covering all the known factors involved in the problem. Part of the desired information may be already available. It is important not merely to add to this mass of data, but to make new material comparable and continuous with what is already in existence.

Economic and social phenomena are usually the results of a diversity of factors. Some of these factors, vital to an adequate explanation of the phenomena, may through ignorance be omitted from the inquiry, and an explanation may be sought on the basis of a part of the essential facts. For example, an investigator, seeking to ascertain why school children are retarded in their school work and compelled to fall behind their proper grades, collects the facts concerning the astonishing number of physical defects found by medical inspection. These defects of eyes, ears, teeth, throat and bodily functions occur in larger relative numbers among the retarded than among the other pupils. The inference might be drawn that physical defect is the chief factor in the retardation of school children. The incompleteness of this inquiry consists in the failure to recognize other possible factors and to investigate them with equal care. The fact that many parents are foreign born and, lacking facility in the new language, render little assistance to their children in school work, the differing ages of the pupils on entering school, the frequency of transfer from one school to another, irregularity of attendance, and mental defects are all factors of varying importance. Until the inquiry includes at least the most probable factors, no valid conclusions concerning causation can be drawn, nor can the relative importance of the factors involved be estimated.

Suppose the problem is to find out the typical yearly earnings of the workers in a specific branch of industry. The actual amounts paid to each worker may be ascertained from the pay-sheets for a week or other payroll period. But in many cases the same individual cannot be followed through the payrolls of an entire year because he has shifted from one employer to another. It would be easy to multiply one week's earnings by 52, but this method takes no account of the time during which the employee has no work. If the trade is seasonal in character the choice of a *representative week* is a difficult matter because earnings vary widely in slack and rush seasons. The inquiry must be planned to meet these and many other similar difficulties, or fail in its purpose. Effective planning requires exact knowledge of the trade.

It is necessary, therefore, in the statistical analysis of any problem that the investigator ascertain in advance exactly what are the possible factors which may influence the result. Otherwise, it is impossible to know what material should be gathered, with what degree of accuracy, or how it should be collected.

A WORKING HYPOTHESIS

Is a working hypothesis essential in the planning and carrying out of a statistical investigation? What is a working hypothesis? It is not a preconceived theory or an explanation which the investigator sets out to defend, to the exclusion of evidence that might support other theories or explanations. Such an attitude characterizes the disputations of advocates who leave to their opponents the task of bringing forward the facts on the other side. This is the narrow attitude of the propagandist whose mind is closed to other explanations to which the facts might lead. It is not the point of view of the scientific investigator.

A working hypothesis, in the scientific sense, is a theory or an explanation held after careful canvass of the known facts, in full knowledge of other explanations that have been offered, and with a mind open to a change of view if the facts disclosed by the inquiry warrant a different explanation. It is, therefore, held *with the definite purpose of including in the investigation all available and pertinent data, either to prove or disprove the hypothesis.* An hypothesis of this character is usually desirable and even essential. It gives point to the inquiry, and, if founded on sufficient previous knowledge, guides the lines of the investigation. Without it much useless data may be collected in the hope that nothing essential will be omitted, or important data may be omitted which could have been easily included if the purposes of the inquiry had been more carefully defined. Blind gathering of masses of data does not usually lead to the discovery of unexpected relations between facts or result in new explanations.

A CLEARLY DEFINED STATISTICAL UNIT

Quantitative data, the raw materials to which statistical methods are applied, can be accurately collected only when the unit to be counted, measured, or estimated has been carefully defined. This statistical unit must be clearly understood by the investigator. If more than one person collects the facts no effort should be spared to avoid misunderstanding of what the unit includes, and as little latitude as possible should be left for personal interpretation on the part of the enumerator or the person from whom the information is sought.

The task of definition may appear easy but there are many unexpected difficulties. For example, most States collect the records of industrial accidents but there is little agreement as to what constitutes an accident for purposes of official record. One State requires an employer to report all accidents, slight or serious, causing any loss of time; another State counts the injury for official record only after the injured person has been absent from work on account of the injury for a specified term, as one week. The number of recorded accidents in the two States for the same industry are not comparable from the point of view of relative protection of life and safety. *The unit counted does not mean the same in both States.*

What is a crime and who is a criminal? Efforts are frequently made to measure anti-social conduct by the records of arrests, convictions or number of prisoners in penal and reform institutions. But definitions of crime vary in different communities. By strict enforcement one locality may record many arrests and convictions for a particular offense; while another locality, because of indifference, may show few cases. We cannot compare the two communities by the use of these facts because what is really recorded is not the same in the two localities.

Suppose one desires to investigate wages in a specific branch of industry. Should weekly rates of pay or weekly earnings be ascertained? If the latter, account must be taken of subtractions for loss of time or damage to goods, and of additions in the form of bonuses, commissions or overtime pay. Is it annual earnings which are to be measured? If so even weekly actual earnings are inadequate because of periods of unemployment for various causes. Here the unit of time covered is important, as well as the unit of pay.

On January 1, 1920, thousands of enumerators began to count the population of the United States in the decennial census. Surely counting persons is a simple task. But in the very first household at which the enumerator called was a visitor. Should this person be counted in the place where found or at his usual place of residence? Without careful and uniform instructions to enumerators this individual will be counted twice or missed altogether. There must be no difference in the interpretation of instructions. Where is a student in other than his home town college to be enumerated? Who will record the required data for the head of a household who is absent on a business trip to another city? In the agricultural census the enumerator records facts about farms. Is the truck garden of several acres just outside the city a farm? An owner of a farm, occupied by himself, has several other tracts of land occupied by tenants, or he lives on one tract and cultivates another at some distance. How many farms should be recorded? If an inquiry is being made concerning the number of rooms in city apartments and the average number of persons per room, should the kitchen and bath be counted in the total for each apartment? The agent of the Bureau of Labor Statistics is collecting the retail prices of food commodities. In two different cities the same cut of beef is given a different name.[1] In two different stores in the same city prices are quoted for eggs or butter but the grades are not the same. The resulting price data are not comparable. In estimating the relative size of universities who is to be counted a student? One university may carry on extension work throughout the entire state while another does little extra-mural work. In the former case are all those with whom extension workers come into contact to be counted as students? Is a person taking one course to be counted just the same as a student taking many courses?

It is evident from these illustrations that clear definition of the unit to be counted is absolutely essential. The unit must have the same meaning and scope at different times or in different places between which comparisons are to be made. *Every effort should be made to adopt objective tests in defining the unit and to avoid the influence of personal interpretation and bias.*

METHODS OF GATHERING DATA IN A FIRST-HAND INQUIRY

There are several methods in current use for gathering first-hand statistical data. The method or combination of methods to be used will influence greatly the nature and number of the items included.

1. **Personal inquiry.** This method can best be used in a very intensive study of comparatively few cases where the need of completing the work quickly is not urgent. Le Play, the French professor of metallurgy, spent his summer vacations making detailed studies of family budgets but he could cover only a small number of selected families in this manner. The personal inquiry brings to its aid the enthusiasm and insight of one who is deeply interested in the problem — one who, it is presumed, understands the limitations and difficulties of the material, and who, if he be scientific in spirit, is vitally concerned with the accuracy of the results.

[1] For pictures showing beef cuts in various cities, New York City, Chicago and New Orleans, see Bulletin 315, United States Bureau of Labor Statistics, pp. 69–73.

The difficulty of training helpers, who have neither the initial interest in the problem nor the same motive to produce reliable results, is avoided.

On the other hand, the number of cases which can be covered is so small that there is grave danger of errors in conclusions drawn from too narrow a range of facts. Added to this is the possibility that the "personal equation" may even unconsciously introduce a bias in the selection of the cases. The results may tend to show what the investigator wishes to show.

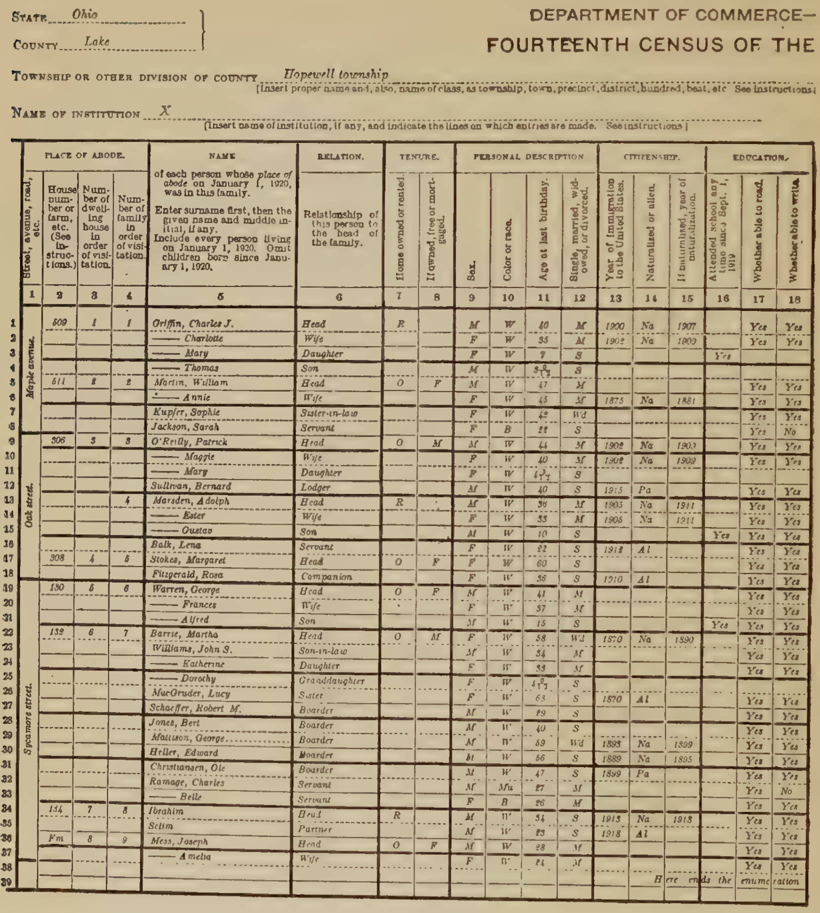
2. **Enumerators or special agents.** In this method the schedule of questions or of items to be covered is presented in person to those who are expected to give the information, or is filled in by special agents from some authoritative source, as payrolls. The Federal Census of Population is taken by this method. To cover the entire population of the United States in about two weeks many thousands of trained enumerators are employed. The enumerator, guided by detailed and standardized instructions, aids the individual in answering correctly questions which may need explanation or interpretation. The questions or items are printed on a card or sheet and the answers are written in the blank spaces, usually by the enumerator or agent. (See pages 376–77.)

This method of gathering data allows a more extensive inquiry to be made, but its expensiveness often prevents its use in a private investigation. It is, however, the most reliable method for extensive investigations. In order to secure uniform results, honest and intelligent persons must be selected for the work, careful instructions and training must be given them, and conferences must be held during the initial stages to settle doubtful questions. Different interpretations of the same question can often be prevented by the clearness of the question itself and the explicit nature of the instructions.

3. **Questionnaires — Schedules filled out by informants.** The questions may be sent by mail or otherwise placed in the hands of the person giving the information. The answers are given without direct supervision or assistance, which differs entirely from the methods previously described. This method permits an extensive inquiry with much less expense than the enumerator or special agent method, because a very large number of cases may be covered almost simultaneously by the use of the mails.

There are many difficulties in the use of this method and in general the results are likely to be unreliable. As a rule, a large proportion of those to whom schedules are sent have little or no interest in and a very limited understanding of the problem. Some are suspicious of the purpose of the inquiry, or have reasons for not stating certain facts. Busy persons are

[ILLUSTRATIVE EXAMPLE OF MANNER



DEPARTMENT OF COMMERCE—

FOURTEENTH CENSUS OF THE

STATE *Ohio*
COUNTY *Lake*

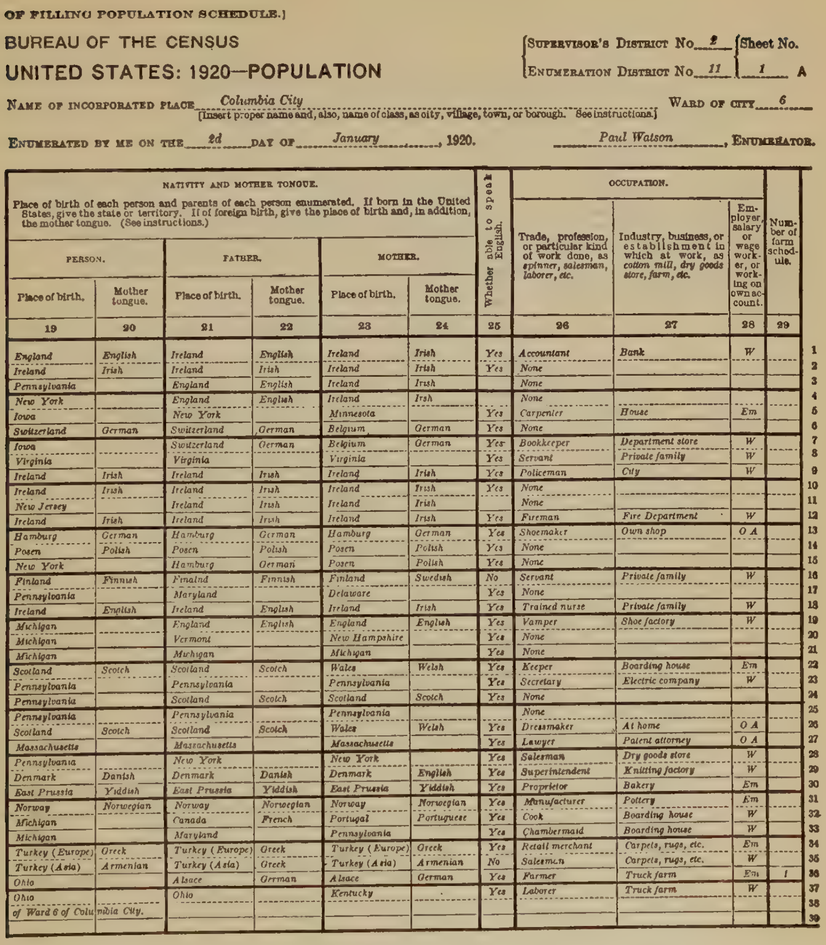
TOWNSHIP OR OTHER DIVISION OF COUNTY *Hopewell township*
[Insert proper name and, also, name of class, as township, town, precinct, district, hundred, beat, etc See instructions]

NAME OF INSTITUTION *X*
[Insert name of institution, if any, and indicate the lines on which entries are made. See instructions]

	PLACE OF ABODE.				NAME	RELATION.	TENURE.		PERSONAL DESCRIPTION				CITIZENSHIP.			EDUCATION.		
	Street, avenue, road, etc.	House number or farm, etc. (See instructions.)	Number of dwelling house in order of visitation.	Number of family in order of visitation.	of each person whose *place of abode* on January 1, 1920, was in this family. Enter surname first, then the given name and middle initial, if any. Include every person living on January 1, 1920. Omit children born since January 1, 1920.	Relationship of this person to the head of the family.	Home owned or rented.	If owned, free or mortgaged.	Sex.	Color or race.	Age at last birthday.	Single, married, widowed, or divorced.	Year of immigration to the United States.	Naturalized or alien.	If naturalized, year of naturalization.	Attended school any time since Sept. 1, 1919	Whether able to read.	Whether able to write.
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	*Maple avenue.*	*509*	*1*	*1*	*Griffin, Charles J.*	*Head*	*R*		*M*	*W*	*40*	*M*	*1900*	*Na*	*1907*		*Yes*	*Yes*
2					——— *Charlotte*	*Wife*			*F*	*W*	*35*	*M*	*1902*	*Na*	*1909*		*Yes*	*Yes*
3					——— *Mary*	*Daughter*			*F*	*W*	*7*	*S*				*Yes*		
4					——— *Thomas*	*Son*			*M*	*W*	$3\frac{0}{12}$	*S*						
5		*511*	*2*	*2*	*Martin, William*	*Head*	*O*	*F*	*M*	*W*	*47*	*M*					*Yes*	*Yes*
6					——— *Annie*	*Wife*			*F*	*W*	*45*	*M*	*1875*	*Na*	*1881*		*Yes*	*Yes*
7					*Kupfer, Sophie*	*Sister-in-law*			*F*	*W*	*42*	*Wd*					*Yes*	*Yes*
8					*Jackson, Sarah*	*Servant*			*F*	*B*	*22*	*S*					*Yes*	*No*
9	*Oak street.*	*306*	*3*	*3*	*O'Reilly, Patrick*	*Head*	*O*	*M*	*M*	*W*	*44*	*M*	*1902*	*Na*	*1903*		*Yes*	*Yes*
10					——— *Maggie*	*Wife*			*F*	*W*	*40*	*M*	*1902*	*Na*	*1909*		*Yes*	*Yes*
11					——— *Mary*	*Daughter*			*F*	*W*	$4\frac{3}{12}$	*S*						
12					*Sullivan, Bernard*	*Lodger*			*M*	*W*	*40*	*S*	*1915*	*Pa*			*Yes*	*Yes*
13				*4*	*Marsden, Adolph*	*Head*	*R*		*M*	*W*	*36*	*M*	*1905*	*Na*	*1911*		*Yes*	*Yes*
14					——— *Ester*	*Wife*			*F*	*W*	*33*	*M*	*1905*	*Na*	*1911*		*Yes*	*Yes*
15					——— *Gustav*	*Son*			*M*	*W*	*10*	*S*				*Yes*	*Yes*	*Yes*
16					*Balk, Lena*	*Servant*			*F*	*W*	*22*	*S*	*1912*	*Al*			*Yes*	*Yes*
17		*308*	*4*	*5*	*Stokes, Margaret*	*Head*	*O*	*F*	*F*	*W*	*60*	*S*					*Yes*	*Yes*
18					*Fitzgerald, Rosa*	*Companion*			*F*	*W*	*36*	*S*	*1910*	*Al*			*Yes*	*Yes*
19	*Sycamore street.*	*130*	*5*	*6*	*Warren, George*	*Head*	*O*	*F*	*M*	*W*	*41*	*M*					*Yes*	*Yes*
20					——— *Frances*	*Wife*			*F*	*W*	*37*	*M*					*Yes*	*Yes*
21					——— *Alfred*	*Son*			*M*	*W*	*15*	*S*				*Yes*	*Yes*	*Yes*
22		*132*	*6*	*7*	*Barrie, Martha*	*Head*	*O*	*M*	*F*	*W*	*58*	*Wd*	*1870*	*Na*	*1890*		*Yes*	*Yes*
23					*Williams, John S.*	*Son-in-law*			*M*	*W*	*34*	*M*					*Yes*	*Yes*
24					——— *Katherine*	*Daughter*			*F*	*W*	*33*	*M*					*Yes*	*Yes*
25					——— *Dorothy*	*Granddaughter*			*F*	*W*	$4\frac{0}{12}$	*S*						
26					*MacGruder, Lucy*	*Sister*			*F*	*W*	*63*	*S*	*1870*	*Al*			*Yes*	*Yes*
27					*Schaeffer, Robert M.*	*Boarder*			*M*	*W*	*29*	*S*					*Yes*	*Yes*
28					*Jones, Bert*	*Boarder*			*M*	*W*	*40*	*S*					*Yes*	*Yes*
29					*Mattison, George*	*Boarder*			*M*	*W*	*69*	*Wd*	*1893*	*Na*	*1899*		*Yes*	*Yes*
30					*Heller, Edward*	*Boarder*			*M*	*W*	*66*	*S*	*1889*	*Na*	*1895*		*Yes*	*Yes*
31					*Christiansen, Ole*	*Boarder*			*M*	*W*	*47*	*S*	*1899*	*Pa*			*Yes*	*Yes*
32					*Ramage, Charles*	*Servant*			*M*	*Mu*	*27*	*M*					*Yes*	*No*
33					——— *Belle*	*Servant*			*F*	*B*	*26*	*M*					*Yes*	*Yes*
34		*134*	*7*	*8*	*Ibrahim*	*Head*	*R*		*M*	*W*	*34*	*S*	*1913*	*Na*	*1918*		*Yes*	*Yes*
35					*Selim*	*Partner*			*M*	*W*	*23*	*S*	*1918*	*Al*			*Yes*	*Yes*
36		*Fm*	*8*	*9*	*Mess, Joseph*	*Head*	*O*	*F*	*M*	*W*	*28*	*M*					*Yes*	*Yes*
37					——— *Amelia*	*Wife*			*F*	*W*	*24*	*M*					*Yes*	*Yes*
38															*Here ends*	*the*	*enume*	*ration*
39																		

more or less annoyed by the request for data which interrupts their routine work. Many schedules, therefore, are not returned at all or are imperfectly filled out. Sometimes those returned are representative of a selected group, and not of the entire field. The government, in case the inquiry is official, may impose penalties for failure to reply or for misstatements. Private investigators using this method must rely upon interest and persuasion.

In this type of investigation it is especially important to submit as few questions or items as possible and to make them very simple, definite and easy to answer. Every additional item increases the danger that the schedule will be consigned to the waste basket or will be returned imper-



OF FILLING POPULATION SCHEDULE.]

BUREAU OF THE CENSUS

UNITED STATES: 1920—POPULATION

SUPERVISOR'S DISTRICT No. 2 | Sheet No.
ENUMERATION DISTRICT No. 11 | 1 A

NAME OF INCORPORATED PLACE Columbia City [Insert proper name and, also, name of class, as city, village, town, or borough. See instructions.] WARD OF CITY 6

ENUMERATED BY ME ON THE 2d DAY OF January, 1920. Paul Watson, ENUMERATOR.

NATIVITY AND MOTHER TONGUE. Place of birth of each person and parents of each person enumerated. If born in the United States, give the state or territory. If of foreign birth, give the place of birth and, in addition, the mother tongue. (See instructions.)						Whether able to speak English.	OCCUPATION.			Number of farm schedule.	
PERSON.		FATHER.		MOTHER.			Trade, profession, or particular kind of work done, as *spinner, salesman, laborer, etc.*	Industry, business, or establishment in which at work, as *cotton mill, dry goods store, farm, etc.*	Employer, salary or wage worker, or working on own account.		
Place of birth.	Mother tongue.	Place of birth.	Mother tongue.	Place of birth.	Mother tongue.						
19	20	21	22	23	24	25	26	27	28	29	
England	English	Ireland	English	Ireland	Irish	Yes	Accountant	Bank	W		1
Ireland	Irish	Ireland	Irish	Ireland	Irish	Yes	None				2
Pennsylvania		England	English	Ireland	Irish		None				3
New York		England	English	Ireland	Irsh		None				4
Iowa		New York		Minnesota		Yes	Carpenter	House	Em		5
Switzerland	German	Switzerland	German	Belgium	German	Yes	None				6
Iowa		Switzerland	German	Belgium	German	Yes	Bookkeeper	Department store	W		7
Virginia		Virginia		Virginia		Yes	Servant	Private family	W		8
Ireland	Irish	Ireland	Irish	Ireland	Irish	Yes	Policeman	City	W		9
Ireland	Irish	Ireland	Irish	Ireland	Irish	Yes	None				10
New Jersey		Ireland	Irish	Ireland	Irish		None				11
Ireland	Irish	Ireland	Irish	Ireland	Irish	Yes	Fireman	Fire Department	W		12
Hamburg	German	Hamburg	German	Hamburg	German	Yes	Shoemaker	Own shop	O A		13
Posen	Polish	Posen	Polish	Posen	Polish	Yes	None				14
New York		Hamburg	German	Posen	Polish	Yes	None				15
Finland	Finnish	Finalnd	Finnish	Finland	Swedish	No	Servant	Private family	W		16
Pennsylvania		Maryland		Delaware		Yes	None				17
Ireland	English	Ireland	English	Ireland	Irish	Yes	Trained nurse	Private family	W		18
Michigan		England	English	England	English	Yes	Vamper	Shoe factory	W		19
Michigan		Vermont		New Hampshire		Yes	None				20
Michigan		Michigan		Michigan		Yes	None				21
Scotland	Scotch	Scotland	Scotch	Wales	Welsh	Yes	Keeper	Boarding house	Em		22
Pennsylvania		Pennsylvania		Pennsylvania		Yes	Secretary	Electric company	W		23
Pennsylvania		Scotland	Scotch	Scotland	Scotch	Yes	None				24
Pennsylvania		Pennsylvania		Pennsylvania			None				25
Scotland	Scotch	Scotland	Scotch	Wales	Welsh	Yes	Dressmaker	At home	O A		26
Massachusetts		Massachusetts		Massachusetts		Yes	Lawyer	Patent attorney	O A		27
Pennsylvania		New York		New York		Yes	Salesman	Dry goods store	W		28
Denmark	Danish	Denmark	Danish	Denmark	English	Yes	Superintendent	Knitting factory	W		29
East Prussia	Yiddish	East Prussia	Yiddish	East Prussia	Yiddish	Yes	Proprietor	Bakery	Em		30
Norway	Norwegian	Norway	Norwegian	Norway	Norwegian	Yes	Manufacturer	Pottery	Em		31
Michigan		Canada	French	Portugal	Portuguese	Yes	Cook	Boarding house	W		32
Michigan		Maryland		Pennsylvania		Yes	Chambermaid	Boarding house	W		33
Turkey (Europe)	Greek	Turkey (Europe)	Greek	Turkey (Europe)	Greek	Yes	Retail merchant	Carpets, rugs, etc.	Em		34
Turkey (Asia)	Armenian	Turkey (Asia)	Greek	Turkey (Asia)	Armenian	No	Salesman	Carpets, rugs, etc.	W		35
Ohio		Alsace	German	Alsace	German	Yes	Farmer	Truck farm	Em	1	36
Ohio		Ohio		Kentucky		Yes	Laborer	Truck farm	W		37
of Ward 6 of Colu	mbia City.										38
											39

fectly filled out. Clearness is fundamental because there is no one present at the time the answers are given to explain the questions or to check up the replies. Of course, where the appeal for information is made to persons already interested in the problem who, for the most part, desire to further the purposes of the investigation, this method is more likely to yield valuable results.

Government offices in a number of States employ this method to collect data on wages, hours, and other industrial facts from employers who are requested to fill out blanks periodically furnished by the statistical bureaus. Likewise, data on industrial accidents are reported on blanks furnished by State labor departments (see page 379). It is found that where these reports are required regularly at stated periods and checked by an efficient central authority, through correspondence or the visits of special agents, the results steadily become more complete and reliable. This method is useful mainly in gathering current information. When the questions require in a single report data extending over any considerable period the probability of serious errors is increased.

Registration of births and deaths. The facts concerning each birth are filled out by the physician or other responsible person on blank forms which have been standardized for most communities (see page 380). A penalty is usually imposed, although not always enforced, for failure to file this report with the proper authorities. When the record is filed it constitutes *an official registration* of the birth at or near the time of its occurrence. This method of *registration* is contrasted with the recording of data by a *census or enumeration method* at stated or intermittent intervals of time.

The data as to each death in a community are recorded on *a standard death certificate* (see page 381). The form when properly filled out is signed by the attending physician or other responsible person, and is filed with the local health authorities in exchange for a burial permit. The certificates of births and deaths, together with reports of sickness, furnish the fundamental data for vital statistics which are essential to effective health administration and the development of sanitary science.

4. Method of estimates. Frequently facts cannot be counted or measured and must be estimated. For example, the United States Department of Agriculture has long used this method in its crop reporting service. Correspondents living in the different localities send in estimates at regular intervals concerning the acreage and the condition of the principal crops. These estimates are usually stated as a percentage of the acreage and yield of the previous year or of a normal period. Where only approximate results are required and where exact enumeration or measurement is not possible estimates prove very useful. In fact, it is an important function of statistical method to make possible more accurate estimates. Business men must resort to estimates of future requirements in labor, equipment and materials and their probable costs. The accuracy of these estimates is checked constantly by the records of current business as they accumulate. The entire system of budgeting in private business and in government departments is based upon more or less scientific estimates.

It is clear that the schedule of questions or items cannot be formulated intelligently until the method of gathering the data has been decided upon. A wider scope of inquiry, more complex questions and a

STANDARD FORM FOR ACCIDENT REPORTS[a]

First Report of Accident to Employee

[To be filled out and sent in within forty-eight hours of the accident.]

---

1. Employer.

a. Employer's name ....

b. Office address: Street and No. ....; city or village ....

c. Business (goods produced, work done, or kind of trade or transportation) ....

d. Location of plant or place of work where accident occurred, if not at office address: Street and No. ....; city or village ....

e. Name of insurance carrier ....

---

2. Injured person.

a. Date on which accident occurred ....

b. Working hours per day ....; c. Working days per week ....

d. Piece or time worker? ....; e. Wages or average earnings per day ....; per week ....

f. Name ....; address ....

g. Sex ....; h. Age ....

i. Occupation when injured ....; in what department or branch of work? ....; was this regular occupation? ....; if not, state regular occupation ....

---

3. Cause of injury.

a. Describe in full how accident happened ....
....
....

b. Name of machine, tool, or appliance in connection with which accident occurred ....; by what kind of power driven? ....; hand feed or mechanical feed? ....; part on which accident occurred ....

---

4. Nature and extent of injury.

a. State exactly part of person injured and nature of injury ....
....

b. Did injury cause loss of any member or part of a member? If so, describe exactly ....

c. Has injured person returned to work? ....; if so, give date and hour ....

d. Date disability began ....

---

5. Medical care.

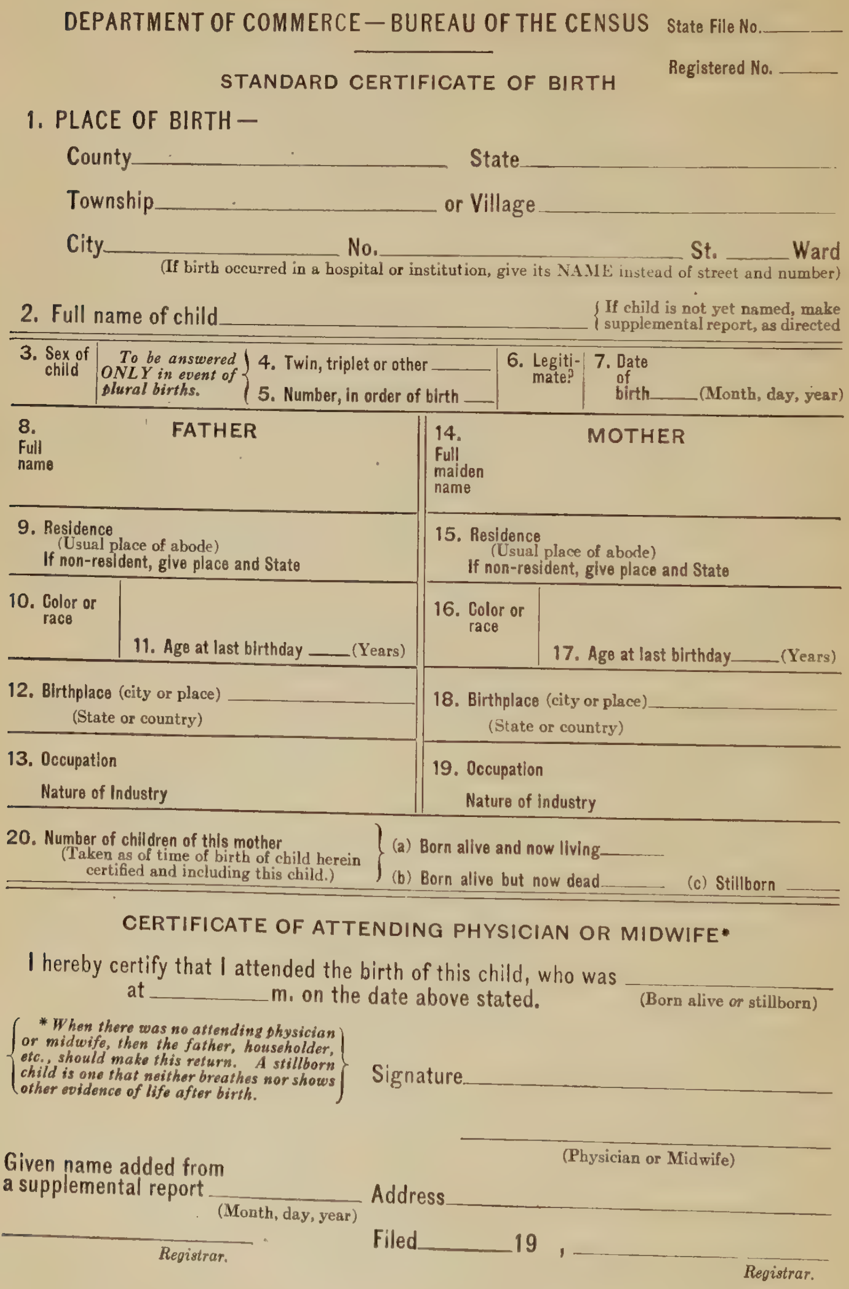
a. Attending physician; name and address ....
....

b. Hospital; name and address ....
....

Date of report ....; made out by ....

---

[a] Bulletin 276, United States Bureau of Labor Statistics, p. 21.



DEPARTMENT OF COMMERCE—BUREAU OF THE CENSUS State File No.______

Registered No. ______

STANDARD CERTIFICATE OF BIRTH

1. PLACE OF BIRTH—

County________________ State________________

Township________________ or Village________________

City____________ No.________________ St. ______Ward

(If birth occurred in a hospital or institution, give its NAME instead of street and number)

2. Full name of child________________________ { If child is not yet named, make supplemental report, as directed

3. Sex of child	*To be answered ONLY in event of plural births.*	4. Twin, triplet or other ______ 5. Number, in order of birth ______	6. Legitimate?	7. Date of birth______(Month, day, year)

8. FATHER Full name	14. MOTHER Full maiden name
9. Residence (Usual place of abode) If non-resident, give place and State	15. Residence (Usual place of abode) If non-resident, give place and State
10. Color or race 11. Age at last birthday ______(Years)	16. Color or race 17. Age at last birthday______(Years)
12. Birthplace (city or place) ____________ (State or country)	18. Birthplace (city or place)____________ (State or country)
13. Occupation Nature of Industry	19. Occupation Nature of industry

20. Number of children of this mother (Taken as of time of birth of child herein certified and including this child.) } (a) Born alive and now living______ (b) Born alive but now dead______ (c) Stillborn ______

CERTIFICATE OF ATTENDING PHYSICIAN OR MIDWIFE*

I hereby certify that I attended the birth of this child, who was ____________ (Born alive *or* stillborn)
at __________m. on the date above stated.

{ * *When there was no attending physician or midwife, then the father, householder, etc., should make this return. A stillborn child is one that neither breathes nor shows other evidence of life after birth.* }

Signature________________________

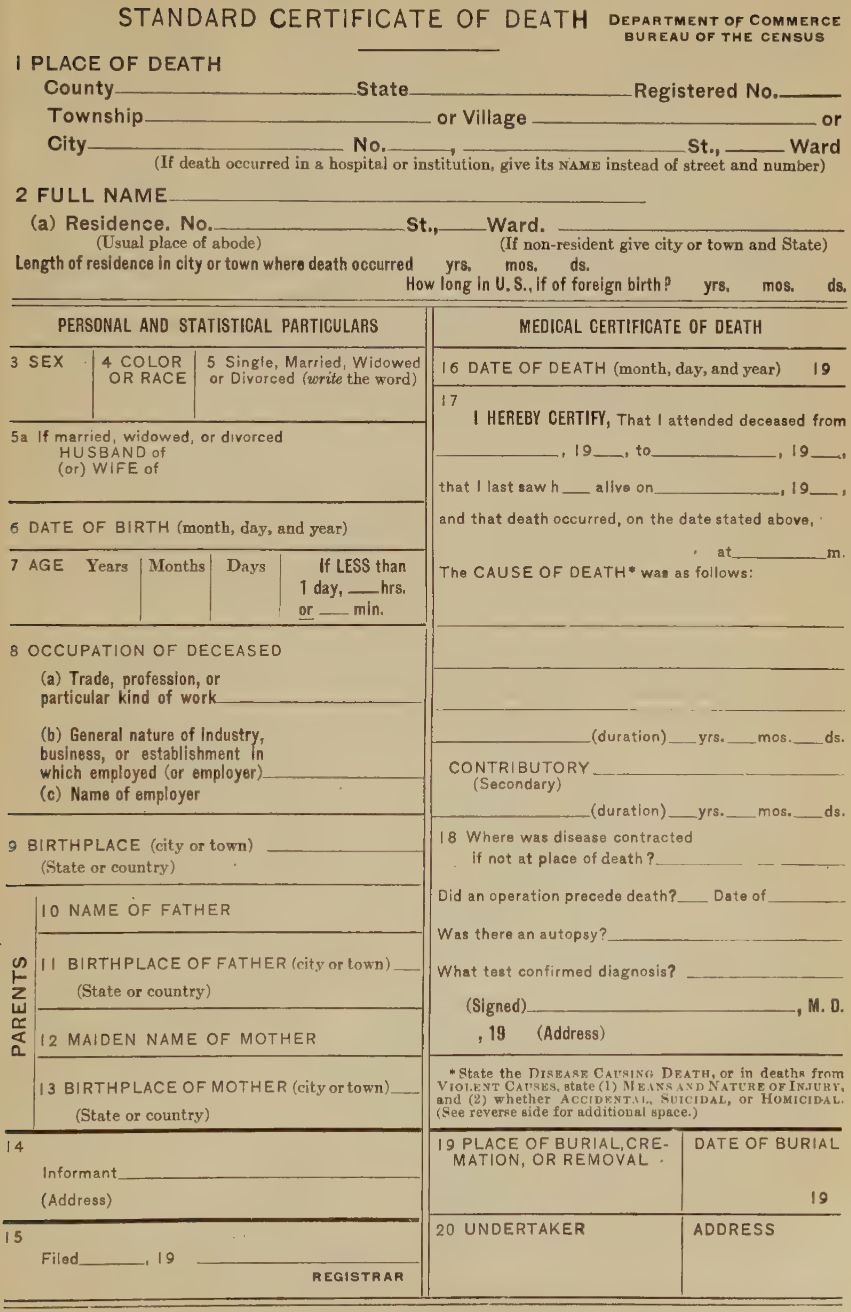
____
(Physician or Midwife)

Given name added from a supplemental report ____________
(Month, day, year)

Address________________________

____
Registrar.

Filed________19 , ________________
Registrar.



STANDARD CERTIFICATE OF DEATH

DEPARTMENT OF COMMERCE
BUREAU OF THE CENSUS

1 PLACE OF DEATH

County________________State________________Registered No.______

Township________________ or Village ________________ or

City________________ No.______, ________________St., ______ Ward

(If death occurred in a hospital or institution, give its NAME instead of street and number)

2 FULL NAME________________________________

(a) Residence. No.________________St.,______Ward. ________________
(Usual place of abode) (If non-resident give city or town and State)

Length of residence in city or town where death occurred yrs. mos. ds.
How long in U. S., if of foreign birth? yrs. mos. ds.

PERSONAL AND STATISTICAL PARTICULARS

3 SEX | 4 COLOR OR RACE | 5 Single, Married, Widowed or Divorced (*write* the word)

5a If married, widowed, or divorced
HUSBAND of
(or) WIFE of

6 DATE OF BIRTH (month, day, and year)

7 AGE Years | Months | Days | If LESS than 1 day, ___hrs. or ___ min.

8 OCCUPATION OF DECEASED
(a) Trade, profession, or particular kind of work________________
(b) General nature of industry, business, or establishment in which employed (or employer)________________
(c) Name of employer

9 BIRTHPLACE (city or town) ________________
(State or country)

PARENTS

10 NAME OF FATHER

11 BIRTHPLACE OF FATHER (city or town)______
(State or country)

12 MAIDEN NAME OF MOTHER

13 BIRTHPLACE OF MOTHER (city or town)______
(State or country)

14
Informant________________
(Address)

15
Filed______, 19 ________________
REGISTRAR

MEDICAL CERTIFICATE OF DEATH

16 DATE OF DEATH (month, day, and year) 19

17
I HEREBY CERTIFY, That I attended deceased from
____________, 19___, to____________, 19___,
that I last saw h___ alive on____________, 19___,
and that death occurred, on the date stated above,
at____________m.
The CAUSE OF DEATH* was as follows:

____

____

____

____________(duration)___yrs.___mos.___ds.

CONTRIBUTORY ________________
(Secondary)

____________(duration)___yrs.___mos.___ds.

18 Where was disease contracted
if not at place of death?____________

Did an operation precede death?___ Date of______

Was there an autopsy?____________

What test confirmed diagnosis? ____________

(Signed)________________, M. D.
, 19 (Address)

*State the DISEASE CAUSING DEATH, or in deaths from VIOLENT CAUSES, state (1) MEANS AND NATURE OF INJURY, and (2) whether ACCIDENTAL, SUICIDAL, or HOMICIDAL. (See reverse side for additional space.)

19 PLACE OF BURIAL, CREMATION, OR REMOVAL | DATE OF BURIAL 19

20 UNDERTAKER | ADDRESS

greater number of them, and questions of a more personal character may be used if the investigator in person, or assisted by enumerators or agents with specific instructions, gathers the facts. *A combination of the methods described above may be employed in the same investigation.*

A REPRESENTATIVE INVESTIGATION

A *complete induction* results from a record of all the cases within the field of investigation about which information is desired and to which the conclusions apply. For example, in 1909–10 a serious strike occurred at the Bethlehem Steel Works and the United States Bureau of Labor Statistics undertook to make a report on the conditions of work in that plant in reference to wages and hours. The special agents of the Bureau consulted the employers' records and tabulated the wages and hours of work of more than 9000 men — the entire labor force at that time. It was a complete induction. The conclusions applied to this one company only and could not be regarded as typical of the conditions in the steel industry generally without first proving that the Bethlehem Steel plant was representative of the whole industry. Sometimes this type of report is called a *monograph,* meaning an intensive complete study covering usually a very restricted range of cases.

In contrast to the complete induction is the *representative investigation.* The field may be large and time and funds may be limited. Besides, for the purpose in view it may not be necessary to gather data concerning every case. The attempt is made to conduct the investigation in such manner that *conclusions may be drawn which apply to the entire field although only a part of the cases have been actually examined in detail. A picture in miniature of a larger whole is sought.* The cases included in the actual inquiry must be typical and must represent fairly the other cases which are not included. If these other cases should all be included it ought not to change the conclusions, provided our method is valid.

In most investigations of any size, especially those conducted by private means, the representative method has to be used to save time and expense, if for no other reason. The Federal Census Bureau, at the decennial periods, covers the entire population with a detailed schedule of questions. On the other hand, the Federal Bureau of Labor Statistics, for the current studies of wages in specific employments, selects typical establishments in various localities covering all the important factors which affect wages in that trade. The facts are taken from payrolls by special agents or are reported on forms prepared by the Bureau and the results are checked. From these samples conclusions are drawn concerning wage conditions in the entire trade.

Which of these methods will be followed in any particular investigation must be decided at the outset and the plans completed in accordance with this decision.

Sampling in a representative investigation. The purpose of a representative investigation is to secure from the samples chosen a true picture on a small scale of a larger whole. Since the balance and perspective will be correct only on condition that the samples actually investigated are truly representative of the entire number of cases, *the manner of selecting the samples is of fundamental importance to the success of the inquiry.*

In the buying and selling of commodities sampling is a familiar practice. The quality of a carload of oranges is judged by opening a few boxes taken from different parts of the shipment. In determining the grade of wheat, samples are examined from various parts of the entire mass of grain, and for this purpose a special instrument has been devised. The retail dealer frequently places the finest fruit on the top of the box but the wary buyer requires him to shake up the berries from the bottom to test the uniformity of their quality. This procedure gives all grades of quality a fair chance to be included in the test samples.

The effect of allowing the sewage of New York City to flow into the surrounding waters is tested by an examination of many small samples of river and harbor water taken from different areas and at varying depths. It is necessary to examine only a very small proportion of the water in the harbor in order to make possible definite and accurate conclusions.

Review of certain aspects of sampling. *In Chapter XI it was shown that averages and other measures calculated from samples are themselves variables.* They differ more or less from the values which would be obtained if all the cases in the field investigated were included. For example, successive samples of 500 each are selected from a group of 10,000 adult men of native birth, and the heights of each sample are averaged. The averages of the samples differ from each other and each may differ from the average height of the entire 10,000. It may be demonstrated that these measures computed from samples approach the values obtained from the whole population as the size of the sample is increased, not directly with the increase in the number of cases but in proportion to the square root of the number of cases in the sample. Therefore, the number of items included becomes an important factor in determining the *adequacy of the sample.* When the sample is of a given size, the expected deviation of the calculated values because of chance variation due to sampling, may be described in terms of probability. In this manner the reliability of a given measure computed from a sample is described in Chapter XI. *In that discussion an unbiased selection of the sample was assumed.*

Furthermore, in homogeneous data there is less tendency to vary than in heterogeneous material, regardless of the size of the sample. Therefore, the measure of variability in the samples becomes another fundamental factor. It follows that in the type of material where the individual measurements are closely massed together and where the range of variation is narrow, not so many cases are required in the sample to render it adequate in reflecting the characteristics of the entire population.[1] For example, in the coal industry so large a sample would not be required to represent adequately the various factors which affect wages as would be necessary in the steel industry where the range of wages paid from lowest to highest is much wider and where the kinds of work and degrees of skill are more varied.

It is perfectly possible to have the size of the sample adequate and the *probable error* (for explanation of probable error, refer to Chapter XI) small, and yet have it fail to represent fairly the larger population from which it has been selected. Suppose, for illustration, that an investigator wishes to show the relation of wage changes in specific industries in the United States during the World War to the movement of prices and cost of living. He collects his wage data, for the most part, from establishments which are non-union, or open shop. In 1914 these plants paid lower wages than the union plants. Furthermore, during the war, wages in the non-union plants, on the average, rose more rapidly than wages in the union shops, because in the former the level was low to start with, the competition for labor was keen, and there existed no agreements on wage rates to retard the movement. The facts, therefore, seem to show that the upward movement of wages during the period equaled or exceeded the changes in prices. The objection may be made that, while this apparently is true for a part of the laborers, the results do not reflect accurately the entire wage situation. A selected group has been used. *The sample is adequate in size but not representative.*

How then, can the representative character of the sample be secured? This is, for the most part, not a mathematical problem but a logical one. It depends upon an intelligent analysis and classification of the factors to be investigated in the entire field covered. *It is of fundamental importance to avoid bias in the selection of the sample.*

The meaning of random selection. The most important principle in

[1] The term *population* is used in the technical sense of the entire number of cases in the field under investigation.

sampling is to proceed in such manner that every case of the entire population covered in the investigation has, as nearly as possible, a fair opportunity to be included in the sample, even though not actually chosen for examination. The method has been illustrated already from the field of buying and selling. If significant differences exist which are likely to affect the final results these differences should be known beforehand in order that a method may be devised to distribute the sample so as to include them.

In economic and social investigations the procedure of sampling becomes more difficult because it is less possible to use the mechanical methods of securing an unbiased selection of cases. Many different factors enter into most of these problems. The units are human beings who differ from each other in many respects, as age, sex, nationality, occupation, degrees of skill or intelligence and amount of income. Therefore, the groups and subgroups, to be represented fairly and in proper proportion in a sample, are very numerous.

Random selection of cases is to be distinguished from careless selection. The latter leads to serious bias without the knowledge of the investigator. He may approach the paysheets of business establishments, for example, with no bias as to what he desires to show from the records. He may be ignorant, however, of the varying size of plants in the industry, of the proportions of skilled and unskilled workers, and of the variety of processes to be represented. Any sample of wage-earners made up under these conditions could scarcely be expected to give a true picture of the wage situation of the entire industry.

An intelligent classification of the factors affecting the wage problem is a preliminary step in obtaining a representative sample. This requires complete information about the particular industry under investigation and discriminating judgment in the definition of groups and sub-groups to be covered. The process of grouping the workers establishes the requisite degree of likeness within each group. Then any cases chosen from a particular group will reflect its characteristics and represent it in the entire number of workers investigated. Care should be taken that the smallest sub-group in the sample has sufficient cases to represent adequately the variations in wages which occur within that class.

It is important that the number of cases chosen to represent any group should bear about the same proportion to the entire sample as the total workers in that group bear to the entire number in the industry. By this procedure not only is each factor affecting wages given a fair representation but random choice of cases within each sub-group gives each case in the entire population a fair chance to be included. In short, care must be exercised to give to each group proper weight in its influence upon the entire sample.

Since adequate classifications are essential in obtaining a representative sample it is urged that the reader review the discussion of this subject in Chapter IV.

The natural inclination is to collect data which are easily accessible. On this account a sample may become a selected group, more restricted than the field originally planned for investigation. For example, it is desired to describe a cross-section showing at a given time the proportion of physical defects among the children registered in the eight elementary grades of a city's public schools. The samples in the different schools are chosen from among those present on a certain day and the examiners do not include those absent on that day. The results are representative of the children present but not of all registered children. Especially if the weather is bad on that day, it is certain that a larger proportion of physical defects would be found among the absent children.

Suppose it is planned to investigate the cost of living for a family of typical size and composition among factory workers in a certain community. The schedule is very detailed, with scores of items of expenditure, covering food, clothing, rent, furnishings and miscellaneous expenses. The investigator asks for both the quantity and the cost of these items covering the period of a year. It is difficult to find housewives who are able and willing to give the desired information or who will keep account books month by month for the purpose. When such persons are found they are likely to possess more than usual intelligence and thrift and their families are not likely to be representative of the group. In other words, the sample is a restricted group. Conclusions drawn from it must be used with caution and tested by further investigation.

Schedules returned by mail are very likely to represent a group different from the one originally marked out. Thousands of schedules may have been sent out, properly distributed, but only ten per cent are returned. Are the ten returned from each hundred sent really random and free from bias or are they selected? Are there questions in the schedule which restrict the number of the returns and which select the individuals from whom they come? It is the representative character of the sample returned and not of the original mailing list which should concern the investigator. Caution in the use of such returns is always necessary. *Judgment and experience are of more importance than formal rules in guarding against wrong conclusions from unrepresentative samples.*

Practical test for the size of the sample. In a previous section of this chapter it has been explained why the minimum size of the sample need not be the same for every investigation. The number of cases required depends largely upon the homogeneity of the data. Assuming that a method has been devised of securing a sample, the concern of the investigator is to examine enough cases to render the conclusions trustworthy. This may be made a matter of actual experiment with the particular type of data. *A useful rule of procedure is to increase the size of the sample until successive analyses and summaries show sufficiently similar results.* For example, in a wage inquiry, when the sample seems adequate according to the plan mapped out beforehand, the results may be classified and summarized in frequency tables and averages. Then, more cases may be added, chosen in a similar manner, and the data may be summarized again. If the differences in the results are so slight as to be negligible for the purposes of the inquiry, the sample may be regarded as *adequate in size.* The object of sampling is to describe the whole from the actual investigation of certain parts. As long as our conclusions about the whole are changed by the inclusion of more of the parts there can be no possible confidence in the accuracy of the conclusions. *Stability of results is essential.*

This practical test of the size or adequacy does not carry a guarantee that the sample is also representative, because all the data may have a bias which mere number of cases will not eliminate. The error involved may be a constant one. (See discussion in Chapter XI.)

THE SCHEDULE OF QUESTIONS OR ITEMS

Emphasis has been placed already upon the fact that the nature of the questions or items and their number depend to a considerable extent on whether they are presented by enumerators in person, or in some other manner are placed in the hands of informants to be filled out. In formulating questions and deciding upon items in a schedule it is necessary to take into consideration whether a public authority or a private person or agency is conducting the investigation. The former has power to compel answers and to punish evasions. On this account the questions in the official inquiry may be somewhat more personal and detailed and their number may be increased.

General considerations in preparing a schedule. 1. *The schedule is of first importance if accurate and complete data are to be made available.* The success of any inquiry, therefore, depends very largely upon the skill with which the items and questions have been formulated and arranged. It is of little use to apply elaborate methods of analysis to fundamentally defective raw materials unless the errors are known and allowance can be made for them. Sometimes errors are concealed by the very method of handling the data. *More reliable data at the source is the greatest present need in the fields of both official and private statistics.*

2. The formulation of the questions requires a most careful preliminary study of the problem in order to know what to include and what to exclude.

3. Schedule making demands knowledge not alone of human nature in general, but specific knowledge of the groups investigated, their intelligence, prejudices, interests and probable reactions. This is especially important if the schedules are sent out by mail.

Important rules in determining the items or questions. 1. *Accuracy and completeness* of the resulting data are the chief aims in formulating the questions. Not all investigations require the same degree of accuracy in the recorded data, or the same completeness in the answers. Standards for each particular inquiry must be determined and carefully defined beforehand, and then the data must be brought, as nearly as possible, up to these standards. Scientific work in chemistry or physics requires very delicate measurements, perhaps weighing with a hair balance; but it would be quite unnecessary, even absurd without a carefully itemized record which most families do not keep, to state to the cent of accuracy the total annual expenditure for food. It is *spurious accuracy* to state the number of bushels of wheat produced annually in the United States to the last unit, because the amount is estimated and not counted. It gives a false impression of accuracy to work out birth-rates per thousand of the population to the *second decimal place* when only ninety per cent of the actual births are recorded in the given city. In the *Weekly Bulletin* of the New York City Department of Health, December 29, 1923, page 415, the average "persons per house" is stated to two decimal places, 68.96 and 53.76.

2. *The questions should be simple and definite.* Qualifying adjectives and adverbs should be avoided because they increase the opportunity for subjective interpretation of the question. For example, it may be asked, "Have you an A.B. degree?" but not, "Are you *well* educated?" In a schedule on industrial accidents, it would be useless to ask, "Was the injured *experienced* in his work?" It would be proper to ask, "*How long* had the injured worked at the process?"

The question, as a rule, should be so phrased as to require *objective answers* rather than *subjective impressions.* Sometimes investigators are sent out to record such matters as sanitation and cleanliness in order to establish grades for bakeries or eating places, as A, B, C, or Good, Fair, Bad, according to specified conditions. The difficulty is to define each grade so definitely that the same recorded grade shall always mean the same actual condition when different investigators make the observations.

The questions and items should be simple enough to be understood easily by the least intelligent members of the group investigated. Otherwise, the imperfect answers of the ignorant destroy the accuracy and completeness of the results. *Average intelligence is a false standard in planning schedules.*

The questions should be so formulated and the items should be so arranged on the schedule form as to make it easy to record the required data. Wherever possible "yes" or "no" or a simple number should be the answer requested. Frequently, to facilitate the recording of answers, lists of items are printed on the schedule, to be checked by the informant opposite the proper items. Simple examples are, male, female, for sex; or, married, single, widowed, divorced, for marital status, using the check (✓) opposite the item which records the answer. This principle may be employed as far as space will permit. When data are sought on such matters as wages, hours, numbers employed in different processes of manufacturing, and value of raw materials, a tabular form should be drawn up on the schedule itself, with all the required items clearly stated, leaving blank spaces for the proper entries. This plan promotes easy and accurate recording, facilitates handling the data from the schedule, and often saves space on the blank form.

3. *As few questions as possible should be included.* Every added item increases the labor and cost of tabulation. The investigator should assure himself that the results will compensate for the effort and expense. This means that questions should be formulated with a definite idea of *how the answers will be utilized.* Limitations of time and funds will control the utilization of answers beyond a certain limit. A trial schedule sometimes reveals the need for further items but it also offers an opportunity to select and eliminate questions. The temptation is to include some item because it may *possibly* be useful. In a schedule sent out for voluntary answers every added item increases the chances of losing the entire schedule in the waste basket. Even in the case of the Federal Census, where enumerators and special agents are used, there is great need of simplifying the schedules. Enumerators and informants are wearied by long lists of items and the accuracy of the results is impaired.

4. *The questions should be such as can be answered truthfully and without bias.* If suspicion or resentment is aroused by a question it is difficult to secure accurate information, not only on the specific question but on the others as well. Questions concerning property and income are often

resented as inquisitorial, or arouse the suspicion that the results will be used for purposes of taxation. Even before the prohibition amendment became law, questions concerning the use of intoxicants were not truthfully answered. For example, a workman, after having been interviewed as to his use of leisure time and as to how many glasses of beer he consumed during the day, was overheard to remark to a companion, "This fellow asks us for information we would not tell our own wives." In formulating questions the investigator must not lose sight of the peculiar traditions, prejudices, and backgrounds of the persons to whom they are to be submitted.

It is usually important to make clear the purpose of an inquiry, and it is frequently necessary to convince the informants that their names are not to be used in connection with the answers. In other words, the names must be kept in confidence while the data obtained are used for statistical purposes. Often schedules are numbered and a perforated slip is attached bearing the same number. The name of the informant, if required at all, is written on this slip which is detached as soon as the schedule is returned to the central office, and is kept only for identification purposes, in case answers must be verified or additions made.

What the lawyer calls *leading questions*, and questions which may be advantageous or disadvantageous if answered in a certain way, are to be avoided. A question which checks another may be included, as, for example, *age* and *date of birth*. The instructions to the enumerator may require that he quiz the informant before recording the final answers to certain types of questions.

5. *Inquiries should cover just the information desired with as little chance as possible for two interpretations.* This point has been emphasized under the definition of the unit and the importance of making questions simple and definite. For example, if the inquiry calls for the number of adult workers in a factory a definite age limit must be stated.

6. *The form of the schedule is important.* The size varies widely according to specific needs. Some schedules are printed on cards and others on sheets of flexible paper, or even on folders with more than one sheet. For example, the Federal Census of Population is taken on a schedule form 16 × 23 inches in size. Both sides of the sheet are used and the required data for 100 persons may be recorded on a single sheet. A detailed family budget inquiry, such as the Federal Bureau of Labor Statistics uses, includes almost 500 items and covers several pages. Wherever possible a schedule should be of convenient size *to carry or send without folding*. It should also conform in size to some one of the

standard filing devices, even if it is in booklet form. These requirements are met very well by a 5 × 8, or $8\frac{1}{2}$ × 11 inch size.

When a schedule is circulated by mail, it is often desirable to make the form exactly like a personal letter, typing, folding, signing and addressing it accordingly, in order to increase the probability of its return.

The quality of the card or sheet should be such as to permit the use of ink, and sufficiently durable to stand handling and sorting many times without replacement. For sorting into groups in hand tabulation, cards are much more convenient and durable than flexible sheets. *Different colors* are used to distinguish classes of schedules, as an aid to the eye in sorting and filing. On the face of the schedule, careful *attention to the spacing* and to the *kind of type* used in printing the headings promotes both ease and accuracy in recording and tabulating the data. After placing the necessary information at the top of the schedule to identify the specific inquiry described below, and having left space for the date and the name of the enumerator or agent handling the schedule, the other items should be arranged in some logical order on the card. The main headings and key items may be emphasized by bold type. These should not be crowded and related items should be localized in well-defined areas of the schedule form in order to facilitate recording the data and tabulating the results. *It is desirable to number or letter each item in the schedule in order to render identification easy and to facilitate tabulation.*

Finally, in estimating the importance of schedule making it should be remembered that *the schedule is a means of scientific observation.* It enables the investigator to make the record of facts in an *objective manner* and to use such records made by different observers of the same or closely similar phenomena. In other words, it standardizes the records and assists the investigator in avoiding personal bias, which enters easily into observation and memory. Detailed records on a schedule, emphasizing one factor at a time, may be made to present a fairly complete picture of conditions under investigation.

Before the final schedule is circulated a trial form should be multigraphed and a certain number filled out under the same conditions as will prevail in the actual investigation. After a study of the results obtained by the trial schedules, errors and omissions can be corrected and the accuracy and completeness of the data finally collected will be greatly improved. This is especially necessary in a new field of inquiry where the past experience of investigators is not available.

THE USE OF DATA GATHERED BY OTHERS

The person interested in statistics frequently uses data collected by others instead of gathering them at first hand. He may depend upon existing sources entirely for his raw materials, or he may supplement available facts by others which are considered essential for the purpose. *He is a consumer of statistics as well as a producer.* In any case, he must judge the data collected or compiled by others in respect to accuracy and availability for his purposes.

Primary and secondary sources. The person who uses statistics should distinguish carefully the sources for which data have been gathered at first hand, from the sources for which data have been transcribed or compiled from the original sources. The population volumes of the Federal Census are examples of the first; the *World Almanac* and the *Statistical Abstract of the United States*, both annual volumes, are examples of the second. The former are *primary sources* and the latter are *secondary sources.* It makes no difference whether a private or a public authority is responsible for the original inquiry. When the responsibility for gathering the original data and for their promulgation is undivided, whether in published form or in other available forms, *the source may be called primary.* But when the authority for the data as promulgated is different from that which controlled the collection of the facts at first hand, *the source containing such data may be called secondary.*

This is a useful distinction for the research worker to keep in mind, because the relative reliability of the two sources as above defined is likely to be different. Transcribing quantitative data from original sources will produce errors not entirely eliminated even by the most careful checking. Besides, compiling statistics from a variety of sources carries the risk that data which are not really comparable will be assembled in the secondary source and treated as if they were comparable, without adequate explanation of their limitations. Furthermore, statistics, adequate for the purposes of the original investigation, may be introduced in the secondary source in relations which do not reflect the true situation. For example, a cost of living investigation, made in approved scientific manner for large cities, may be cited to indicate the inadequacy of the wages paid in coal mining districts. The data available are not directly applicable without modification in the settlement of a wage dispute in the coal industry. It would require additional field work on the cost of living in the coal districts, if the issue is to be decided on this basis.

It is necessary, therefore, for the scientific worker to scrutinize secondary sources very closely. Their reliability for research work can be determined only by reference to the primary source, which should be cited in notes or bibliography. This will enable any one who so desires to make himself responsible for the facts by reference to the original source. Discrepancies appear in different secondary sources which must be settled from the original source. For example, the population of a State in 1920 is given at different figures by the *World Almanac* and the *Statistical Abstract.* The ultimate source is the Census Bureau at Washington, which both collects these facts originally and is responsible for their publication in the volumes on population issued under its own direction. There is a divided responsibility for the figures from secondary sources, since one agency collects the data and another agency compiles and publishes them.

Guiding principles in the use of sources. No attempt will be made to discuss exhaustively the principles which should guide the student in the use of sources, but some of the most important considerations are suggested in the following paragraphs.

It is never safe to accept published quantitative data at their face value. They may not be adequate for the purpose in view, and we must go behind the published figures in order to discover their possible limitations. Some who use published statistics seize upon them eagerly in the spirit of an advocate or debater. If the facts do not exactly answer the purpose, effort is made to press them into service, in the absence of more adequate information, without examination or explanation of their shortcomings. Then some one who knows the conditions of the original inquiry points out how the facts have been used for a different purpose than that for which they were originally collected; how the conclusions have been extended over a wider scope than the data warrant; or how certain new factors in the situation have been neglected entirely.

What has just been said presupposes that the original data have been carefully and accurately gathered and have been honestly and intelligently classified and used. This is by no means always the case. The consumer of published facts, in order to protect himself from the mistakes and deceptions of so-called scientific investigators, must patiently seek to find out for himself how they were collected and what they really mean.

In using available sources it is essential for the research worker to take the same point of view as if he were about to collect the necessary information by a first-hand inquiry. In consequence of this attitude, he will take steps to find out whether the existing data conform to his requirements. In the original inquiry *how was the unit defined?* Does this definition agree with the one in the mind of the present investigator? As pointed out in a previous section, the unit is more or less arbitrarily defined in any investigation, *and the apparent meaning of a quantity is not always its real significance.* A first-class report of an original inquiry should furnish a copy of the schedule used, together with a careful explanation of the units enumerated, measured or estimated. The method of selecting the samples and the procedure in collecting the data should be explained. From such a report the person who wishes to use the results of the inquiry will be able to satisfy himself as to whether the information is representative, unbiased and reliable, and whether it is adequate for his purposes.

When items which show variation have been grouped, is the classification too crude for the purpose in view? Should there be more age classes in the population, or more degrees of skill distinguished among the wage-earners? Where comparisons have been made are they valid? At least it should be decided whether the measurements or values compared have been determined according to uniform definitions and instructions.

In examining a statistical report it is desirable to think over deliberately what facts should have been collected for the purpose and how the work should have been done. As the results of the inquiry are reviewed it will appear how far the problem has been understood, whether the various factors have been included, and how adequate the data are for reaching the conclusions stated. Yet, many use published statistics without this critical attitude, *especially if the results happen to be in accord with their ideas.*

SUMMARY

In this chapter an attempt has been made to set forth the logical steps in a first-hand statistical investigation and to suggest some of the fundamental principles which should govern in the use of existing sources of information. Time and errors are saved and the probability of useful results is increased by a thorough preliminary study of the problem and the field of investigation. The attempt on the part of one who plans an inquiry to visualize the procedure from the beginning to the final presentation of results will insure that attention is given at the proper time to the essentials discussed in the preceding pages — the time and money available, the scope of the inquiry, the definition of the units, the availability of the facts needed, the questions relative to the use of samples and their distribution, the method of gathering the data, the planning of tabular forms for utilizing the facts collected, the number and content of the questions or items, the form of the blank, and the revision of the tentative schedule from actual trial results.

Emphasis is again placed upon the importance of making tentative forms for the tabulation of the results as a guide to the formulation of a sound *schedule.* The specific headings of the proposed tables will indicate exactly what facts are needed and will suggest a clear logical arrangement, which will facilitate later tabulation of the actual data. A trial schedule will allow correction and revision of the items and give some idea of the reaction of individuals to the inquiry.

Finally, if existing sources are used, the scientific worker should adopt the *critical attitude* which will lead him to review the work of others by the same standards that he would use if collecting the facts for himself by original inquiry. The content of this chapter is addressed, therefore, both to the investigator who collects his data by means of a first-hand investigation and to him who assembles his material from existing sources.

READINGS

King, W. I., *Elements of Statistical Method*, part II. (Chapter 8 on approximation and accuracy in King's text should be read in connection with Chapter 1 on measurement in Weld's *Theory of Errors and Least Squares.*)

Bailey, W. B., and Cummings, John, *Statistics*, chaps. 1–4.

Secrist, Horace, *An Introduction to Statistical Methods*, chaps. 2 and 3.

—— ——, *Readings and Problems in Statistical Methods*, chaps. 2, 3, and 4. (Valuable illustrations of procedure in collecting data and in sampling.)

Bowley, A. L., *Elements of Statistics*, 4th ed., part I, chaps. 2 and 3.

—— ——, *An Elementary Manual of Statistics*, part I, chaps. 7 and 8. (Chapter 7 is an excellent statement in brief scope of the problem of securing a representative sample.)

Jerome, Harry, *Statistical Method*, chap. 2. (The sampling process.)

Chapin, F. S., *Fieldwork and Social Research*, chaps. 2, 5, 6, 7, and 8. (Chapter 5 treats sampling, with illustrations.)

Whipple, G. C., *Vital Statistics*, 2d ed., chap. 4. (Enumeration and registration methods.)

Pearl, Raymond, *Medical Biometry and Statistics*, chap. 3. (The raw data of biostatistics.)

Riegel, Robert, *Elements of Business Statistics*, chaps. 6 and 7.

REFERENCES

Rossiter, W. S., *A Century of Population Growth*, Bureau of the Census, 1909. (Discussion of census methods.)

Willcox, W. F., "Development of the American Census Office Since 1890," *Political Science Quarterly*, September, 1914.

—— ——, "The Statistical Work of the United States Government," *Quarterly Publication of the American Statistical Association*, March, 1915.

Bowley, A. L., *The Measurement of Social Phenomena.*

Falk, I. S., *The Principles of Vital Statistics.* (A simple presentation of results in this field.)

Methods of Procuring and Computing Statistical Information of the Bureau of Labor Statistics. Bulletin 326, United States Bureau of Labor Statistics, Washington, 1923. (This bulletin describes briefly the methods of collecting data on wages, cost of living, prices, accidents and employment.)

Standardization of Industrial Accident Statistics, Bulletin 276, United States Bureau of

Labor Statistics, Washington, 1920. (Standard accident reporting forms and standard table forms for analysis and publication in reports.)

Carr, Elma B., "Cost of Living Statistics of the United States Bureau of Labor Statistics and the National Industrial Conference Board," *Journal of the American Statistical Association*, December, 1924. (A comparison of methods and results.)

Income in the United States, 1909–1919, Volumes I and II, National Bureau of Economic Research, New York City. (Volume II gives excellent illustrations of scientific estimates.)

The History of Statistics, edited by John Koren, published for the American Statistical Association by The Macmillan Company.

EXAMPLES OF SAMPLING

Hilton, John, "Enquiry by Sample: An Experiment and Its Results" (with discussion by Bowley, Edgeworth, Greenwood and Yule), *Journal of the Royal Statistical Society*, July, 1924. (Investigation of unemployment. Samples of different sizes and the accuracy of the results.)

Course of Employment in New York State from 1904 to 1916, Department of Labor, State of New York, Special Bulletin 85, July, 1917. (Appendix, pp. 37–43, gives good example of procedure in obtaining representative data.)

Unemployment in New York City, Bulletin 172, United States Bureau of Labor Statistics, Washington, 1915, pp. 6–8.

Bowley, A. L., and Burnett-Hurst, A. R., *Livelihood and Poverty*, London, 1915, pp. 11–18 and chap. 6. (A social study of four English cities of moderate size.)

Drachsler, Julius, *Intermarriage in New York City*, Studies in History, Economics and Public Law, Columbia University, vol. 94, no. 2, 1921, pp. 19–23.

Burdge, H. G., *Our Boys, A Study of the 245,000 Sixteen, Seventeen and Eighteen Year Old Employed Boys of the State of New York*, Military Training Commission, Bureau of Vocational Training, Albany, 1921, chap. 1.

Mark, Mary L., and Croxton, F. E., "Unemployment Survey in Columbus," *Monthly Labor Review*, April, 1922, pp. 14–23. (Data gathered by students in Ohio State University.)

Pearl, Raymond, *Medical Biometry and Statistics*, chap. 12, pp. 255–62.

For the more technical and mathematical treatment of the theory of sampling the reader should consult the readings and references at the close of Chapter XI.

[Details as to publisher and date of publication of the readings and references given at the close of each chapter are found in the alphabetical list in Appendix A.]

CHAPTER XV

PRESENTING STATISTICAL DATA IN TABLES

EDITING RETURNS ON SCHEDULES

Before the data returned on the schedules are classified and summarized for statistical purposes, the answers are scrutinized to detect, wherever possible, erroneous, inconsistent, or incomplete statements. This procedure is called editing, and its purpose is to secure greater accuracy, to detect inconsistent answers, to facilitate classification and tabulation by making the returns more uniform, and to note omissions in the schedule which may be supplied either by the editor or by referring the item again to the original informant.

In editing it must be remembered that all replies on the schedule are equally original, and that "the only evidence competent to justify the revision of a reply is the evidence presented in the other replies."[1] It may be necessary finally to classify inconsistent replies in a "no report" or "unknown" group. The editor is never justified in erasure and his revisions should appear in distinctive ink or pencil on the face of the schedule. For example, both age and date of birth have been stated and the two facts do not agree. From other evidence on the schedule the editor may be justified in accepting the date of birth as the more probably correct and he may enter the age accordingly. On the other hand, a person may be described as married, head of a family and employed, at age eight years. Possibly a figure two has been omitted but the editor, in the absence of a date of birth, will probably classify the age as "unknown." Sometimes a grand total is made up of sub-totals of separate groups of items. The sub-totals may each check accurately with the sum of their individual items but the sum of the sub-totals may not equal the grand total. This sort of check is more likely to be required in tables than in schedules.

Sometimes different terms or statements appear in the schedules describing exactly the same fact, as occupation or cause of death. The editor seeks to establish uniformity by writing in blue pencil the name of the occupation or the number of the cause of death, as found in the international list of causes, now widely accepted. This procedure makes it easy to tabulate the answers in their proper group.

Often the schedule is returned with blank spaces which should have

[1] Bailey and Cummings: *Statistics*, p. 18.

been filled. Sometimes the editor can supply the proper answers from other replies on the same schedule. If the questions left unanswered are very important the schedule is returned for more complete data. If the physician states two causes of death without indicating their relative importance he is requested to revise the statement. Where answers cannot be supplied the editor must classify omissions as "unknown." *As a rule the editor accepts as final the replies entered upon the schedule, but if there is a presumption of error he attempts to verify the answer.*

The function of tabulation. *The raw material of a statistical inquiry is to be found on the schedules,* sometimes hundreds of separate cards or sheets, sometimes millions, as in the case of the Federal Population Census. The data have not been analyzed, classified, or combined. No one knows how many persons there are in a given geographic location, of each nationality, age, sex, and occupation; or what proportions these classes form of the totals. When the schedules in a wage investigation have been returned, no one knows how many receive a certain specified wage or what proportion this number forms of the total workers in a particular trade, or how wages differ with the years of experience of the worker, or what is the typical weekly wage for a specific kind of work. All of this information and more, the schedules will reveal when properly handled. *It is the function of tabulation to classify, arrange and summarize in easily accessible form the answers to the questions with which the inquiry is concerned.*

The table is the means of presenting more or less detailed statistical information in a compact form, and usually in such manner as to emphasize comparisons and to show relations. The knowledge of how to plan and to construct tables and skill in doing it constitute a very important part of the equipment of a statistician. *Tabulation is not merely a mechanical process* of ruling the page into certain spacings with proper headings; combining individual items from the schedules into totals; computing percentages or rates; and, then, arranging these in the appropriate spaces on the page and checking the correctness of the results. In a modern statistical organization mechanical devices perform most of the drudgery and leave the energy of the statistician free to plan the ways and means of presenting the data in new and illuminating relations, and to interpret the results.

Preliminary to intelligent tabulation of statistical data is an adequate analysis of the factors in the problem which is being investigated. What divisions and subdivisions of geographic location, of age and sex, of race and nationality, or of occupation and skill are required to set forth the data clearly for the purpose in view? *Intelligent tabulation begins with* *the making of the schedule* and the visualization of the scheme for utilizing the results of the inquiry. Why cumber the schedule with questions which have no apparent utility? If a schedule on industrial accidents includes inquiries concerning the experience of the injured person the presumption is that the investigation aims to show the relation, if any exists, between the amount of experience and the frequency of accidents. If the occupation of the deceased is recorded on a certificate of death, the object is to classify deaths by the occupation of the deceased and, in this manner, to show differences between occupations from the point of view of the hazard to the worker. *The planning of tables to present specific data in relation to other data must begin with the formulation of the questions themselves,* as explained in the preceding chapter.

Useful tabulation presupposes logical classifications. This subject was discussed in Chapter IV. The table furnishes the mechanical form for the entry of summaries, rates, and percentages under the categories and subdivisions which are considered essential.

ACCESSIBILITY OF THE ORIGINAL SCHEDULES

Rarely does it happen that the data on the original schedules are easily available to those who use the tables, sometimes because of the confidential character of the information, for example, the reports from individual business concerns, sometimes because of the unwieldy bulk of the schedules and the time and expense involved in a second handling.

The Federal Census schedules on population are now available to localities, at their own expense, for more specific and detailed tabulations of data having particular local interest but not of sufficient general interest to warrant the expenditure of national funds for tabulation and publication in the decennial volumes. For example, the population of a large city may be tabulated by units of area smaller than the ward or assembly district, at present employed by the Census Bureau.[1] This smaller area of tabulation permits death-rates to be calculated per thousand of the population in districts more homogeneous than the political subdivision from the point of view of housing, sanitation, nationality, and other health conditions. If only the larger areas are used for such rates the favorable health conditions in one part of the area are combined with the bad conditions in other sections. The resulting death and sickness rates represent neither the good nor the bad, and do not furnish an accurate measure of relative health conditions in different sections of the city for the guidance of the local health administration.

[1] For an example of this tabulation by small areas, see *Greater New York*, 1920, published by the New York City 1920 Census Committee.

The very fact that the original data are not usually available places a greater responsibility upon the statistician who plans the tables and arranges and summarizes the results. The collection of the facts may be well done in every respect, and yet the results may be spoiled by careless or unintelligent tabulation. *The consumer of statistics is at the mercy of those who plan and execute the tabulations.* It is one of the chief functions of the statistician to set forth the collected data in such manner as to throw into relief the interesting and informing relations which the tables may be made to reveal truthfully under his skillful handling. Quantitative data would cease to be "dry bones" if properly presented, and would take on a lively interest not centered in the figures themselves but in the impressions they convey and the inferences made possible. The figures now describe changes in business activity and the standard of living; the cost of industry in dangers to life and health; the progress of sanitary science and preventive medicine. Quantities thus represent happenings in a kinetic society and the variety of relations and changes which are constantly going on about us.

THE TAKING OF THE DATA FROM THE SCHEDULES

Having decided upon the various kinds and groups of data to be tabulated it remains to sort the facts from each schedule into these groups and to arrive at totals. This may be done by *hand* or by *mechanical methods.*

1. *Hand methods.* If hand methods are to be used in tabulation the schedule should be on cards of durable material and of standard size, as suggested in the preceding chapter, a separate card for each person or unit investigated. These can be sorted easily into any required categories and sub-groups, a separate sorting being necessary for each category desired. For example, weekly earnings, hours, nature of work are recorded on separate cards for each employee in a factory. These may be sorted first into departments, or if it seems desirable, different colored cards may be used in the first place to distinguish different departments. A second sorting may be made according to the kinds of work done in each department. Each of these sub-groups should be sorted to separate men and women, and then each sub-group may be distributed by a final sorting into classes according to the size of the earnings. Totals in any group or sub-group are recorded by simply counting and checking the cards. Care should be taken to make the sum of all sub-totals equal to the grand total which was subdivided. In order to make every sorting a unit which will account for all the cards, it is usually necessary to have a group designated as "unknown" or "not reported" into which may be sorted the cards on which the item being classified is not given or is doubtful.

If the original schedule is too complicated to be printed on a card of size convenient for sorting, or if for any other reason a card cannot be used for recording, it is often possible and may save time to select the most important items from the original schedule and to enter them on such a card. In doing this, numbers and letters corresponding to the numbered items on the original schedule, abbreviations and code numbers may be used to lighten the clerical work of transcribing and to simplify the sorting. For example, many occupations may be represented on the different schedules. Each type of work to be classified may be given a code number, and only the number need be entered on the card used for sorting, provided the same number always means the same type of work. Code numbers are sorted more easily than named occupations. The same procedure may be followed for other items on the schedule.

The tabulation work sheet is a hand method, alternative to the method of sorting just described. First, the scheme of classification, either for all items on the schedule, or more often for only a part of them, is arranged in box headings at the top of the work sheet, with column rulings for each division and heavy or double rulings to distinguish categories which are subdivided, as sex, nativity, age, etc. The horizontal lines of the work sheet are numbered on the left, corresponding to numbers on each schedule passed in review. This number permits checking the accuracy of any entry from the schedule. The following simple form illustrates the method:

	Sex		Age		Marital Status				Nativity			
Schedule Number	M.	F.	Under 16 years	16 years and over	S	M	W	D	NB	NBFP	FB	Occupation Code Number
1	√			√	√				√			8
2		√	√		√					√		10
3	√			√		√					√	7
4		√		√			√		√			6
5	√		√		√						√	12
6	√			√		√				√		10
7		√		√				√	√			6
8	√			√	√				√			5

All the items included in this classification are taken from each schedule in succession at one handling and are recorded by a check (√) under the proper heading, except occupation which is designated by a code number because of the large number of occupational groups. It is easy to total the checks in each column, and the total checks in any category, as age or nativity, must equal the total schedules covered. Of course, cross-classifications are desired. For example, what are the age and sex groupings of all in the occupation designated ten? This requires grouping all ten's according to the age and sex columns. What proportion in occupation six is under sixteen years, and what proportion is married men? The answer requires grouping all sixes again according to the checks in the sex, age, and marital-status columns.

It simplifies the work sheet first to sort out from the original schedules those in a specific occupation, and then to classify one occupation at a time according to all the required categories, handling the schedules for that occupation only once thereafter. It is desirable to have under each of the main categories, as age, sex, etc., a column for the "unknown" in order to make every category a unit covering all the schedules tabulated. Additional sheets with the same headings may be added as the number of schedules tabulated increases, the column totals being stated at the bottom of each sheet and carried forward to the next sheet. Various modifications of the scheme explained and illustrated for tabulating many items at one handling of the schedules are in current use.

It is clear that this *work-sheet tabulation is much less convenient and flexible and is to be avoided wherever sorting the schedules is possible.* Errors are more likely to occur and it is more difficult to check the accuracy of results.

2. *Mechanical methods.* Where the number of schedules is large and the items very numerous, and where detailed classifications and cross-classifications are required, hand methods become too slow and expensive. Electric machines are employed to sort the data into groups, to count the cards and to add quantities which appear on the schedules. For the use of machine sorting the data are transferred from the original schedules to standard cards by means of punching holes in the card. The area of the card is divided into many columns and rows and each small space is designated by a number. All data on the schedules, not already in the form of a quantity, as occupation, country of birth, name of salesman, customer, town or state, are given code numbers. A specific area of the card is assigned to each category on the schedule and the facts are recorded by holes punched in this area. A "punch card" when prepared represents each schedule or unit. A sample card is shown in Figure 60, with the

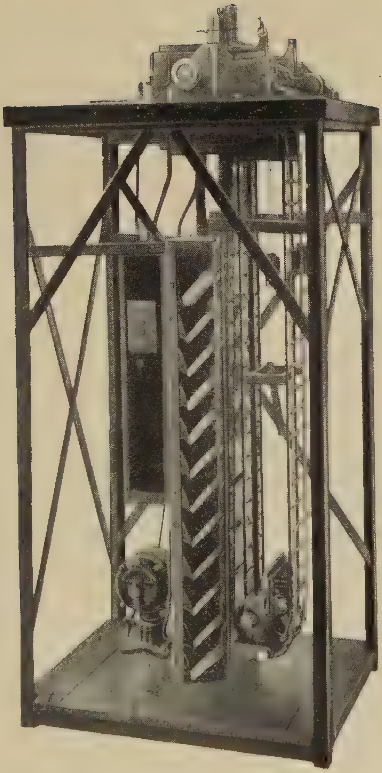


Fig 58. The Electric Sorting Machine

(A mechanical means of classifying data.)

Courtesy of The Tabulating Machine Company.

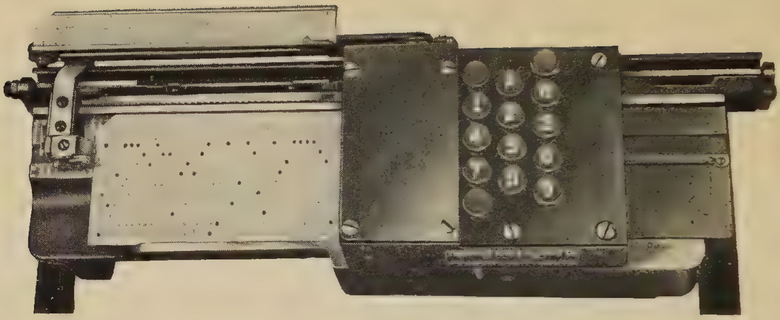


FIG. 59A. THE KEY PUNCH

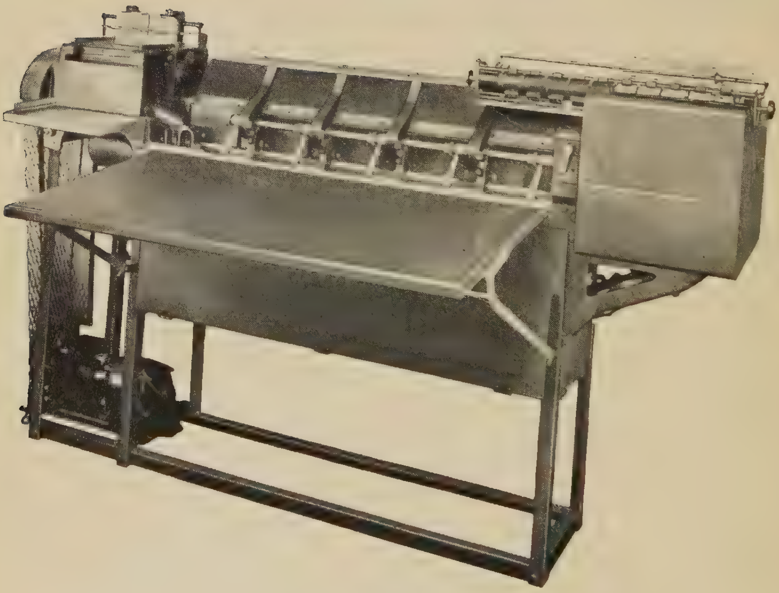


FIG. 59B. THE ELECTRIC TOTAL PRINTING AND LISTING TABULATING MACHINE

(Used to count the cards or to add similar items from many cards, with a printed record of the items and the totals.) *Courtesy of The Tabulating Machine Company.*

holes representing the data recorded concerning the sale of a commodity in a single transaction. Figure 59A represents the Key Punch by means of which the record is transferred to the card, Figure 59B the Electric Tabulator, used to count the cards or to add similar items from many cards, and Figure 58 the Electric Sorter, used to group similar items in required classes.

These cards when punched are verified and are then run through electric sorting machines at the rate of several hundred per minute. The cards fall automatically into boxes which group them according to the required classifications. These sorting and counting machines are used by the Federal Census Bureau and by large corporations. By their use it is possible to secure groupings and correlations of facts which would be impossible by hand methods, due to the time and money required.

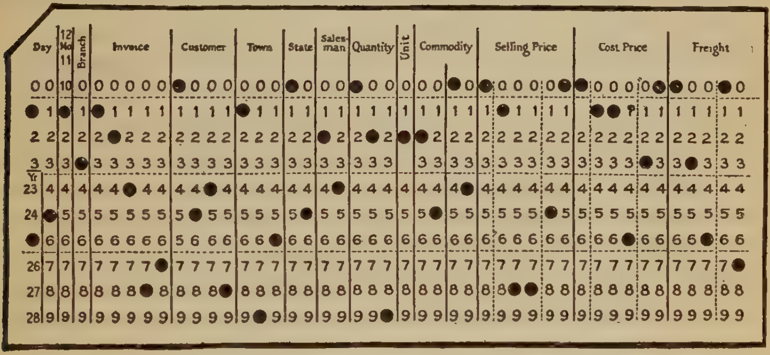


FIG. 60. A SALES CARD PREPARED FOR MECHANICAL TABULATION

(Dots represent holes made with a "key punch" and record data concerning a single transaction.) *Courtesy of The Tabulating Machine Company.*

KINDS OF TABLES

When the classifications have been made and the data have been sorted according to the categoiies and subdivisions required, whether by hand or by machine methods, the results in the form of totals, percentages, averages or rates are finally ready for presentation in statistical tables. If the facts are of wide and varied interest, as in the case of the Federal Census, and are used by many different persons for practical and scientific purposes, they must be presented from many points of view and a great variety of subdivisions is required to meet these different demands. For example, the Census Bureau publishes the age of the population by single years up to 25 and by different age groupings, in order to meet the requirements of those interested in school attendance, mortality statistics, industrial and scientific investigations. To bring together in the most convenient and accessible form these fundamental classes and detailed subdivisions, with the data summarized under each for convenient reference, is the function of *general-purpose or reference tables.*

It is useful to distinguish from general-purpose tables *those designed for a special purpose, which may be called derived or text tables.* These are such as an author would introduce in a book or a magazine article as the basis for his hypothesis or as evidence in support of his argument. The data presented are restricted to certain phases of a problem and the aim is to bring out significant relations in a clear and emphatic manner. For example, the object may be to set forth the differences in age groupings of those engaged in various occupations; or causes of death in relation to occupation; or wages in relation to the length of experience of the worker; or the illiteracy of the population in relation to nationality or to sections of the country. Too many factors and relations must not be crowded together in one such table, at the risk of destroying its emphasis and interfering with the convenience of the reader. Figures must be made inoffensive, if not actually attractive, by clothing them with human interest. Skillful presentation in tables and graphs can accomplish this object.

The general-purpose or reference table has a very different function. It is designed mainly as a source of fundamental data, giving summations of individual items taken from the original schedules and classified under a variety of categories. *Classification conceals the identity of individual units and combines similar cases into groups for statistical purposes.* The general table records these summaries, usually arranged according to some plan which facilitates easy reference, as for example, the population of states in alphabetical order. Many columns and rows and both absolute numbers and percentages are included in a single table, since the primary object is not to show specific relations but rather to serve as a source of information. Good examples will be found in the Federal Census volumes and in other government reports. From these tables may be selected, for further analysis in a series of special purpose tables, the significant factors and relations which require emphasis.

CONSIDERATIONS IN THE CONSTRUCTION OF TABLES

1. *Simplicity of tabular arrangement.* A fundamental problem in presenting the results of an investigation is to decide upon the number of separate tables required. Sometimes funds and publication space are the determining factors. The decision depends upon the details of classification and cross-classification required. The complexity of the

subdivisions and the variety of relations which it is desirable to show may make necessary several tables. For example, the mortality of a city should be shown by geographic areas. No less important are the classifications according to nationality and occupation. In all these analyses cause of death, age, sex and color must be distinguished.

The convenience of those who use the tables should be kept in mind in deciding upon the number of tables and what should be included in each. The table should be made compact and should include as much information as possible in a given space, provided always that clearness is not sacrificed and that the specific object of the presentation is not defeated by the inclusion of details which divert the attention from the main purpose. *A table which is difficult to comprehend will not receive the attention which the facts deserve.* To attempt to show too many distinct factors and relations in the same table often weakens the attention given to any single factor. It will, therefore, often prove more effective to introduce a series of tables, each devoted to a specific object, clearly and simply presented. *The considerations urged above apply mainly to special-purpose rather than to reference tables.*

The *correlation table* is a good example of the special-purpose type. The object is to exhibit the relationship, if any, existing between two series of data, and to emphasize the variations that take place in one series concomitantly or simultaneously with variations in the other series. Both rows and columns of this kind of table are frequency distributions. This arrangement of data is employed in Table 80, page 406, to show the relation between the age and grade of children in the elementary schools. Horizontally in rows the children of a given age are classified according to their proper grades; and, vertically in columns, those in a specific grade are distributed according to ages. Other examples will be found in the chapter on "Correlation," Chapter XII.

2. *Absolute numbers, percentages, and ratios in tables.* In primary reference tables, the chief function of which is to summarize the original tabulations, the absolute numbers should be included in order that the data may be available for a variety of uses and combinations. Percentages and rates may be included to increase the ready utility of the tables. For some purposes the absolute numbers are essential, and for other purposes only percentages of the total or rates per unit of population are significant. *Percentages based upon small absolute numbers are very unreliable.*

When absolute numbers and per cents are given in the same table for several categories occupying a number of columns or rows of the table, the absolute numbers should be given in adjacent columns or rows, and

TABLE 80. GRADES AND AGES OF PUPILS ATTENDING PHILADELPHIA SCHOOLS, DECEMBER 15, 1908[a]

AGES IN YEARS	ELEMENTARY SCHOOLS — GRADES								
	1	2	3	4	5	6	7	8	TOTAL
Under 5.......	5	......	......	......	......	......	......	......	5
5............	1,000	......	......	......	......	......	......	......	1,000
6............	11,841	255	1	1	......	......	......	......	12,098
7............	8,244	5,641	330	3	......	......	......	......	14,218
8............	3,528	7,635	4,518	312	1	......	......	......	15,994
9............	1,137	4,510	5,929	3,186	257	6	......	1	15,026
10............	503	2,463	5,013	4,778	2,660	247	5	......	15,669
11............	250	1,123	3,044	4,352	3,929	2,191	258	12	15,159
12............	143	663	1,853	3,335	3,906	3,538	1,924	219	15,581
13............	63	323	902	1,933	2,950	3,476	2,681	1,619	13,947
14............	19	101	223	541	1,022	1,754	1,953	2,117	7,730
15............	6	12	40	90	225	552	795	1,470	3,190
16............	2	......	5	10	46	133	212	582	990
17............	......	1	......	1	3	25	39	145	214
18............	......	1	......	......	2	5	2	32	42
19............	......	......	......	......	......	......	......	5	5
20............	......	1	......	1	......	......	......	1	3
Total.......	26,741	22,729	21,858	18,543	15,001	11,927	7,869	6,203	130,871
Over age....	5,651	9,198	11,080	10,263	8,154	5,945	3,001	2,235	55,527
Per cent over age.......	*21.1*	*40.4*	*50.6*	*55.3*	*54.3*	*49.0*	*38.1*	*36.0*	*42.4*

[a] *Quarterly Publication of the American Statistical Association*, June, 1911, p. 575.

the corresponding per cents should be placed in another area of the table in adjacent columns or rows for which the titles are repeated. This arrangement facilitates comparisons of either the absolute numbers or the per cents. A simple illustration is presented in the following table form:

NUMBERS AND PER CENTS OF THE TOTAL WHITE POPULATION ACCORDING TO NATIVITY, BY STATES, 1920

STATE	TOTAL	NUMBERS			PER CENTS		
		NB	NBFP	FB	NB	NBFP	FB
1.......							
2.......							
3.......							
4.......							
5.......							

On the other hand, in special-purpose, derived tables, where analysis and inference in the discussion of a particular problem are the important considerations, absolute numbers may only confuse the issue and complicate the table. For example, in the study of family budgets the purpose may be to show what proportions of the total expenditures, in families with different sized incomes, are devoted to food, rent, clothing, fuel and light, furnishings, and miscellaneous items. Obviously, percentages are the only significant figures, as indicated in Table 81, page 419.

In case a relief society wishes to know how much money is required to provide food or clothing for a family of a particular size the absolute numbers in dollars and cents become essential. If we wish to rank the different causes of death in a community according to their numerical importance the actual number of deaths from each cause is required. But if the mortality in different sections of the city having different numbers of people exposed is to be compared, absolute numbers become useless. Only the rate per unit of population exposed (for example, per 1000) is significant. For research it is essential to have somewhere in published form the absolute numbers of deaths classified according to cause, and the population grouped according to all the detailed categories of age, sex, geographic location, nationality and occupation in order to permit a variety of combinations and the computation of specific rates for the various groups.

3. *The use of round numbers.* In accordance with the principle of simplicity, *in derived tables round numbers should be used*, expressed in hundreds, thousands, or millions. In making comparisons and discussing relations attention should be concentrated upon the quantities merely as a basis for deductions. Stating the figures exact to the last unit may increase the difficulty of keeping the various items in mind while adding nothing to the accuracy necessary for comparison. Of course, the rounding of numbers assumes that somewhere, in the source tables, the exact figures are available for reference and checking purposes, and for making new combinations which may then be rounded for use in other special-purpose tables. It should be noted also that in tabular presentation economy of space is usually important. *Rounding figures saves space.*

4. *The general title and the column and row headings. Every table should be as nearly self-explanatory as possible.* Therefore, the main title and the subheadings should convey in clear and unambiguous statements the content of the table. Where it is not possible in descriptive headings to convey the exact scope of the table and its limitations, full explanatory notes should accompany it. When the data are compiled from either primary or secondary sources the references to these should be given

with exactness, both as a defense for the maker of the table and as a convenience to the user. The one who consults tables has a right to be very impatient if he must labor to determine the precise meaning of data presented under indefinite headings or carelessly designated units. It frequently means correspondence with those who are familiar with the original data in the effort to settle uncertain points. For example, an examination of the titles and explanatory notes of the mortality tables published in the annual health reports of some of our large cities does not reveal whether still births have been included or excluded. An otherwise well-constructed table is rendered useless until this information is secured.

The designations for the columns may be termed *box headings* and those of the rows *stub headings.* These must be stated in precise terms, as pounds, inches. Wherever possible the *wording should be printed horizontally* in order to be read in the natural position of the page. Coördinate and subheadings may be indicated by the manner of ruling the box or by indentation of the stub designations.

Each table should be a unit. The classification in any field divides the total into groups more or less detailed, as for example age of the population. But some do not give their age. In order to represent the total cases covered by the inquiry it is necessary in many tables to provide one or more columns or rows with the heading "unknown," or "not specified." Footnotes should be used to explain any peculiarities.

For convenience of reference all tables, except small " text " tables, should *always* be numbered.

5. *The arrangement of the data in columns and rows.* The internal arrangement of a table depends to a considerable extent upon the distinction already emphasized between general, reference uses, and the special-purpose analysis. *Determining the appropriate positions for items in a table involves deciding what shall be placed in columns and in rows, the order of columns and rows, and the conspicuous locations in case emphasis or ready and accurate reference is desired.* Practice varies in these matters, but it will be useful to emphasize certain considerations, keeping in mind the two types of tables.

The general-purpose summary should meet the tests of clearness of statement and convenience and accuracy of reference, but is not limited by the requirement that its entire arrangement shall contribute directly to some specific interpretation. There is a constant tendency to include more categories and subdivisions in a general table, which is likely to expand it beyond the capacity of a single page. Large sheets which must be folded in order to conform to the size of the text page, whether attached or loose, are very objectionable. A table may be spread over two pages facing each other, but even this arrangement is to be avoided if possible. *The ideal is to limit the table to a single page which can be read when held in the same position as the text.* The space limitations of the page often determine which items are assigned to columns and to rows. The ordinary page has much greater capacity vertically. Therefore, the series having the greater number of items is naturally assigned to the columns. Frequently it is necessary to continue the same column headings on several successive pages to complete the table.

In a general table the order of the columns and rows is not so definitely determined as in the case of the special-purpose table. In the former the object is to make the data most generally accessible for reference and transcription. The *alphabetical order,* where appropriate, as in the case of the population of states and cities, clearly fulfils these requirements. Sometimes the *chronological* is the natural order, as the population at successive censuses; or the *geographical location,* as sections of the United States from the Atlantic to the Pacific, or contiguous districts of a city.

On the other hand, the arrangement of categories *according to size* is more likely to be appropriate to a special-purpose table because it commits the analysis to a particular viewpoint, as in the ranking of industries according to the value of their products or the numbers employed. It is the specific purpose of the table that must determine whether the order of the categories shall be according to size, time, or location.

What position should totals occupy in a table? The logical position, followed generally by the accountant, is at the right-hand margin and at the bottom of the table. From this point of view the total combines items distributed along the row or down the column. But in the Federal Census volumes totals are placed at the top and on the left-hand margin of the table, next to the box and stub headings. These are the conspicuous positions and the viewpoint is that of a total distributed into separate items. The makers of these tables justify the practice on grounds of easy and accurate reference, since more people are interested in the totals than in the separate items. This is a valid justification in general-reference tables, but in special-purpose tables as a rule it seems better to follow the logical order of placing totals on the right and at the bottom, where most readers expect to find them. Often in this type of table the totals are little more than checking devices, as in the case of percentages which add to 100. The table is interpreted as a whole because it has a specific purpose. However, in case it is desired to emphasize the totals or any other series of items in a special-purpose table they should occupy the conspicuous positions.

What should be the relative position of earlier and more recent dates in a table? The generally understood position for the earlier date is at the top and on the left of the table. Then movement in time extends from top to bottom and from left to right. In graphic representation this rule is generally accepted. The Census Tables reverse this order, placing the more recent date at the top and on the left. Their reasons are the same as for totals. The recent data are more frequently sought and, therefore, should occupy the conspicuous positions, near the column and row headings, in the interest of both accuracy and convenience. This seems justified for general-reference tables, although there is some danger of errors in transcribing time items located in this reverse order if the entire series is copied at one time. *However, in derived, special-purpose tables the natural order should be followed* of showing progress in time from top to bottom and left to right, in accordance with graphic practice. There is no reason for departing from this rule, for the relation of time items is more significant than the emphasis upon the more recent date or upon convenience of reference. *Effective interpretation of statistical data is promoted by holding to the order which appeals to the readers as most natural.*

As a rule, columns or rows intended for comparison in any table should be placed as near to each other as possible. In preparing a special-purpose table, where specific comparisons are desired, it is found more effective to place the quantities to be compared in columns rather than in rows, because the eye follows up or down a column more easily and accurately than across a row. This is natural since the figures are located one above the other in proper position.

6. *The use of special ruling, spacing, and type in printing.* The object in ruling the table into columns and rows is not alone to guide the eye to the data under a specific heading, but in addition to present at a glance the scheme of classification into main categories and their subdivisions. The proper ruling of the box headings will show coördinate and subordinate classes. Heavy or double lines for columns should be used to separate the main headings, and lighter or single ones for subdivisions. In this manner the complex table may be marked off into areas devoted to particular groups of items which can be located readily and accurately.

In large tables with many rows across the page a blank space left at intervals of five or ten rows aids in locating specific items quickly and accurately. In a wide table having many columns the rows may be numbered at each side of the page with the same numeral in order to aid the eye in locating the desired item from either margin. This device is a substitute for the repetition of the stub headings on the right-hand margin of the table, and has the same purpose as the placing of the numerical scale on both margins of a statistical diagram. *Columns of a table should usually be designated by a number or a letter.* By numbering or lettering both columns and rows reference to any quantity in the table is rendered easy and accurate.

The contrast between *light-face* and *dark-face type* or between *ordinary type* and *italics* should be utilized in printing tables, for the purpose of distinguishing at a glance between absolute numbers and percentages, or between separate items and their totals.

From the above discussion it is clear that a *tentative form for any proposed table should be drawn up in advance.* In this way it is possible to determine experimentally the order and position of columns and rows; the width of columns for the most economical use of available space, some being devoted to large absolute numbers and others to smaller percentage numbers; the best adjustment of the table to the size of the page, or to a double page or to a series of pages; the clearest arrangement of the box and stub headings; and the proper rulings to show clearly main categories and their subdivisions.

IMPORTANCE OF CHECKING ACCURACY IN TABULATION

In all statistical work accuracy is fundamental. A small error in transcribing items, in adding quantities, or in calculating averages, percentages, and rates may not seriously affect the conclusions, but it does always tend to discredit the work. An error in units might have been an error in thousands or in millions. *There is every reason to associate mistakes with careless and incompetent workers.* Confidence in the entire tabulation may be shaken by the discovery of even minor errors.

On this account, checking devices must not be neglected. Each separate item entered in a table should be checked to its source to be sure that the entry is correct. Items in a table with cross classifications, as in a correlation table, should be verified by adding the columns and the rows and then combining these aggregates both horizontally and vertically to form a grand total. If the same grand total results from combining the aggregates in two directions it is an excellent, though not an absolute check on the accuracy of the individual entries (see Table 80). Likewise, when a total is divided into several subdivisions, as for example age groups of the population, the sum of the component parts must be equal to the whole. For this reason it is essential to have a class designated "unreported" or "unknown." Otherwise, the sum of all the groups will not equal the entire population.

When percentages are used to show the component parts of a whole, the separate items should be added to test whether the sum of the parts is equal to 100 per cent. A separate row or column at the bottom or at the right of the table for the actual entry of 100 per cent, the total of a given column or row, serves the double purpose of a check upon the accuracy of the individual entries, and, at the same time, reminds the reader that the figures recorded in the columns or rows are component parts of a whole (see Table 81, page 419). Due to the use of decimals, true to one or more places, the sum of these component percentages frequently amounts to a fraction more or less than 100. Some statisticians feel that, in the interest of mathematical accuracy, the totals should be entered exactly as they are, with the fraction over 100 or less than 100 expressed, and with a note attached to explain the apparent error. The author recognizes that this procedure is mathematically accurate *but logically untrue. The whole can be neither more nor less than* 100, except as the particular mathematical procedure in the use of decimals makes it so. Others make it a practice to adjust the decimals of the separate items at the points where the least error is involved *until the total is exactly* 100, with a note attached explaining the adjustment. The objection to this practice is that the item adjusted by changing the decimal is not consistently accurate as compared with the other items not adjusted. A separate item may be taken from the table and used for a different purpose where there is no test of adding to 100 per cent. In this case it should be accurate. The author believes that this difficulty can be met by leaving the separate decimals mathematically accurate for each item or component part of 100 per cent. Then the total should be stated as 100 per cent even if the parts when added amount to slightly more or less. Such is the present practice of the Bureau of the Census. A note might explain the reason.

Adding machines, such as those commonly used in banks and commercial houses, print a record of each item added, from which the accuracy of the items may be verified by comparison with the original figures. This procedure should never be neglected. If additions are performed without mechanical aid they should be verified by repetition, either by the same person, or, better still, by a different person. Most machines which multiply and divide furnish no written record of the items. Therefore, whether these processes are performed with or without mechanical aid, they should be repeated or checked by the same or a different person.

When "punch cards" are used for mechanical tabulation it is obvious that the accuracy with which these cards are punched from the original schedules is very important. There are several methods of checking the accuracy of the punching before sorting the cards. One method is to read the facts from the punched card while a clerk compares them with the original schedule. Another effective method of detecting errors is to have a duplicate card punched from the same schedule by a different operator. The two cards, supposed to represent the same facts, are compared over a ground glass plate below which is located an electric light. If the holes in both cards are in exactly the same place, as shown clearly by the light, it may be assumed that the facts have been correctly punched. If the duplicate card is of a different color the comparison of the two cards is facilitated. Often two sets of cards are useful for other purposes; for example, one set in the statistical department of a business and the other in the finance department in making up the payroll. It may be inconvenient for both departments to use the same cards. A third method of checking is by a verifying punching machine which is so constructed as to detect errors in the original punching. Unless the accuracy of the original entries on the cards has been guarded by every possible precaution it seems a waste of energy for the statistician to labor over the various combinations and comparisons which are possible. As the analysis proceeds and the interesting interpretations absorb the attention, elaborate statistical methods are likely to obscure these fundamental errors. Sometimes, the operator is paid according to the number of cards punched in a day, which makes some checking device absolutely essential.

EXPLANATION AND INTERPRETATION FROM TABLES

Is it a statistician's business to interpret the results of an investigation? Do the facts speak for themselves when tabulated? Sometimes they do in no uncertain tone, and correctly, if the tables have been carefully planned and annotated. The purpose and content of the accompanying text differ for the two types of tables described in this chapter. There is no uncertainty concerning the special-purpose table. It is presented to emphasize a specific exposition and interpretation. It is definitely related to a background of discussion. The reader is prepared for its introduction and the text sets forth the propositions on which the facts are expected to throw light, their relation to each other, their limitations, and the specific conclusions which the data seem to justify. If relations of cause and effect appear, they are emphasized.

The good judgment of the statistician is required to determine how much weight should be given to the facts resulting from any particular investigation. In some inquiries the data are much more complete and convincing, the conclusions more positive and inclusive than in others. No one should be better qualified to estimate the significance of the data and their limitations than the person or persons who have organized and carried out the details of the investigation, who have studied the problem from every known angle in order to throw new light upon it by the present inquiry, who are familiar with all the difficulties of gathering the material and of putting the results in summary form. *An expert statistician must possess the power to analyze and to interpret results with clearness and understanding.* This ability presupposes a training in the technique of statistical investigation and in the methods of treating and presenting statistical data. In addition, it requires a thorough knowledge of the field in which the statistical expert works and to which he adapts the technical methods required by the nature of the problems to be investigated. *In short, a statistician deserves the title of expert only in the specific fields in which his knowledge is thorough enough to enable him to make productive use of statistical methods.*

The government statistician. In case the data are presented in general-purpose reference tables with adequate notes a minimum of text explanation is required because the purpose is to summarize the facts tabulated from the original schedules and not to emphasize particular interpretations. Many government statistical publications in the United States, local, state, and federal, devote a maximum of space to such tables, setting forth the crude data in a variety of relations and in great detail. Comparatively little space is used to present series of special tables for the purpose of interpreting the crude data. Bulky statistical reports are turned out, with very little attention to the interpretation of the data.

The government statistician is likely to regard it as his function to gather and to publish facts, not to interpret them. It is important to make available the raw material, as indicated in the chapter on index numbers where the need for the periodic publication of original price data was emphasized. Certainly, with the complex schedules of questions at present used in many official inquiries, it would be impossible to reduce the results, in reasonable time or at possible expense, to a form from which specific interpretations could be made. It remains for those interested in making use of the facts to analyze and compare and interpret them for their various purposes.

However, it is fair to raise the question whether our government statistics would not be improved by gathering less facts in bulk and by spending more expert labor and expense in their analysis and interpretation. Planning an inquiry involves the determination of what data are needed and for what purposes. This function of the statistician, when adequately performed, prepares the way for proper classification of the data and for final analysis and interpretation of the results of the investigation. To omit this final step means to stop short of completing the inquiry. To deny this final function to the government statistician takes away an effective incentive to the highest type of professional statistical work. It puts a *premium upon technique and routine* instead of upon the *utility of the results.* To call in an outside expert to do the interpretation means to divide the responsibility for gathering the data and for their interpretation.

Of course an official statistician who does analyze and interpret data must occupy a position free from political influence; he must have a tenure of position dependent upon effective service as an expert — a professional future; and he must have training for his duties in the special fields investigated.

COMPARABILITY OF TABLES

In interpreting the significance of statistical data it is essential to be able to compare one set of facts with another. Therefore, so far as possible, similar data in successive tables, either in the same report or in a series of reports, should be rendered comparable. This is a common failing of state and local statistical reports in the United States. For example, in a certain state, a law is passed requiring more employers to report the facts as to industrial accidents. The official report for the following year may show an increase in the number of industrial accidents when no such increase has really occurred. The report should compare only the industries required to record accidents during the preceding year. In 1898 New York City expanded from the two boroughs, Manhattan and Bronx, to include the present five boroughs of the Greater City. Because of this change of area the mortality rate for New York City in 1899 would not be comparable with that for 1897 unless the 1899 rate were limited to the two original boroughs. At least for a few years following the formation of the Greater City both a rate for Manhattan and Bronx and the entire city should be published for purposes of comparison with the years preceding the consolidation. *New data may be included in a table without impairing its comparability with a preceding table provided the new facts are handled so as to make it possible to separate them from the facts which continue the old series.*

There is little uniformity among the states and localities in the form of the tables presenting similar data. Each state and local statistical office, for the most part, works out its own plan of tabulation. Differences in age grouping, accident classifications, wage classes, and groupings of crimes and criminals appear in the various reports. Some progress in standard classifications and in agreement on the best forms for tabulation has been made in recent years but it is only a beginning.

Order has been brought out of this chaos of incomparable statistics in a number of fields by the current activities of the Federal authorities. For example, facts about coal mining accidents are collected and analyzed by the Bureau of Mines. The Interstate Commerce Commission collects and publishes accident statistics for all railways at quarterly intervals and issues an annual volume on general railway statistics. The Bureau of Labor Statistics gathers data on wages and hours in various industries located in many different states, and on the retail and wholesale prices of a wide range of commodities from many markets. The Division of Vital Statistics of the Census Bureau publishes annually a volume on the mortality statistics of the "registration area" and a volume on birth statistics for a somewhat smaller area. In order to avoid the difficulty of securing comparable data from the states many urge the extension of the statistical work of the federal government. *One thing is certain — that greater uniformity in methods of gathering and presenting facts in various sections of the country is essential in order to render similar data comparable either at the same time or at successive periods.*

SUMMARY

The function and procedure of tabulation have been explained. Two types of tables have been distinguished according to their uses. Certain more or less widely accepted practices in tabular presentation have been discussed. The table should be logically a unit setting forth related data. It should be self-sufficing as far as possible, with a clear explanation of the items included. The general title and the column and row headings should be clear, brief, and self-explanatory. Footnotes must be added if needed to define the units used and to explain any peculiarities of the data. The sources of the facts must be indicated. Coördinate and subordinate headings should be shown by the arrangement of the box and the stub, and by the ruling of the table. Different rulings, spacing and type serve the useful purpose of aiding the eye to locate specific items in the table. For the same purpose columns and rows may be numbered or lettered.

Emphasis has been placed upon special requirements according to the type of table. *Flexibility in planning tables is desirable, and practice can be standardized to advantage only within broad limits.*[1] The general-purpose table is designed primarily to summarize the data from the schedules. Clearness and general accessibility of the facts are fundamental requirements. The assignment of data to columns or rows is frequently beyond the control of the statistician, and is determined by the

[1] Edmund E. Day: "Standardization of the Construction of Statistical Tables," *Quarterly Publication of the American Statistical Association*, March, 1920.

shape of the page and the space requirements. Folders are to be avoided. The order of arrangement of the categories may be alphabetical, chronological, or geographical, but the determining factor should be ready and accurate reference and transcription. Conspicuous locations in the table should be used for items to which reference is likely to be most frequent.

The situation is different in planning a derived or special-purpose table. The object is specific analysis and comparison, and the desire is to enlighten the reader with such exposition and interpretation as the facts and their relations warrant. Simplicity is essential. Round numbers are used. Either percentages or absolute numbers are employed, but rarely both. Arranging figures for comparison vertically in the column positions assists the eye. The logical rather than the most conspicuous positions for totals and for later periods of time are usually best because in derived tables the emphasis is likely to be upon relationships, not upon items. The order of the columns and rows is determined by the specific purpose of the analysis, according to size, progress in time, geographical location, component parts of a whole, etc.

No pains must be spared to check the entries and calculations in a table, and to make the data presented comparable with other similar facts.

READINGS

King, W. I., *Elements of Statistical Method*, chap. 9.
Jerome, Harry, *Statistical Method*, chap. 3. (Note examples, pp. 43–49.)
Secrist, Horace, *An Introduction to Statistical Methods*, chap. 5.
—— ——, *Readings and Problems in Statistical Methods*, chap. 5.
Bailey, W. B., and Cummings, John, *Statistics*, chap. 5.
Bowley, A. L., *Elements of Statistics*, 4th ed., part I, chap. 4.
—— ——, *An Elementary Manual of Statistics*, part I, chap. 6, and part II.
Rugg, H. O., *Statistical Methods Applied to Education*, chap. 3.
Pearl, Raymond, *Medical Biometry and Statistics*, chaps. 4 and 5.
Whipple, G. C., *Vital Statistics*, 2d ed., chap. 2 (pp. 20–22 and 39–57).
Giffen, Robert, *Statistics*, chap. 19.

REFERENCES

Day, E. E., "Standardization of the Construction of Statistical Tables," *Quarterly Publication of the American Statistical Association*, March, 1920.

Watkins, G. P., "Theory of Statistical Tabulation," *Quarterly Publication of the American Statistical Association*, December, 1915.

Standardization of Industrial Accident Statistics, Bulletin 276, United States Bureau of Labor Statistics, Washington, 1920. (Standard definitions and classifications, and standard table forms for analysis and publication.)

Hollerith, H., "An Electric Tabulating System," *School of Mines Quarterly*, Columbia University, April, 1889, vol. x, pp. 238–255. (Describes plans and the first mechanical equipment developed for the 1890 Census.)

"Development of Mechanical Methods," *Proceedings of the Fifteenth International Congress on Hygiene and Demography*, 1912, vol. 6, pp. 73 and 83.

[Details as to publisher and date of publication of the readings and references given at the close of each chapter are found in the alphabetical list in Appendix A.]

CHAPTER XVI

GRAPHIC REPRESENTATION

Within the limits of a single chapter a comprehensive treatment of graphic methods is impossible. Several excellent texts devoted entirely to this subject are now available. Instruction in general statistical methods should include the use of at least one of these treatises, a selected list of which is given at the close of this chapter.

This discussion will deal briefly with the nature and purposes of graphic representation, typical examples of badly constructed diagrams, fundamental rules and cautions in planning graphs or in using those presented by others, and specific types of graphic devices. The difference between scales based upon absolute numbers and ratio or logarithmic scales will be explained and illustrated. Finally, we shall reprint, by permission, the "Preliminary Report, Joint Committee on Standards for Graphic Presentation." This was a committee representative of statistical, economic, engineering, and other scientific associations, brought together under the leadership of Willard C. Brinton of the Society of Mechanical Engineers.

THREE WAYS OF PRESENTING DATA ON FAMILY EXPENDITURES

1. Quantitative data are sometimes presented in the same form as any other descriptive material. For illustration,[1] families having, in 1902, annual incomes under $200 spent, on the average, 16.9 per cent of their total incomes for rent, 8.0 per cent for fuel and light, 50.9 per cent for food, 8.7 per cent for clothing, and 15.6 per cent for other purposes. Families having incomes of $200 and less than $300 spent 18.0 per cent for rent, 7.2 per cent for fuel and light, 47.3 per cent for food, 8.7 per cent for clothing, and 18.8 per cent for other items. The group of families with incomes of $300 and less than $400 spent 18.7 per cent for rent, 7.1 per cent for fuel and light, 48.1 per cent for food, 10.0 per cent for clothing, and 16.1 per cent for other items. The percentages may be stated in this manner for each income group. Since the purpose, however, is to compare the percentages for a specific item, as food, in the different income groups, this form of presentation is confusing and entirely unsatisfactory.

2. By arranging the same data in a table, as is done on page 101 of the original report, the comparisons can be made easily and accurately.

[1] The data are from the *Eighteenth Annual Report of the United States Commissioner of Labor*, 1903, p. 101.

TABLE 81. PER CENT OF EXPENDITURES FOR VARIOUS PURPOSES IN 11,156 NORMAL FAMILIES, BY CLASSIFIED INCOME, UNITED STATES, 1902

CLASSIFIED INCOME (1)	RENT (2)	FUEL AND LIGHT (3)	FOOD (4)	CLOTHING (5)	SUNDRIES (6)	TOTAL (7)
Under $200	16.9	8.0	50.9	8.7	15.6	100.0[a]
200 and under 300	18.0	7.2	47.3	8.7	18.8	100.0
300 " " 400	18.7	7.1	48.1	10.0	16.1	100.0
400 " " 500	18.6	6.7	46.9	11.4	16.5	100.0[a]
500 " " 600	18.4	6.2	46.2	12.0	17.2	100.0
600 " " 700	18.5	5.8	43.5	12.9	19.4	100.0[a]
700 " " 800	18.2	5.3	41.4	13.5	21.6	100.0
800 " " 900	17.1	5.0	41.4	13.6	23.0	100.0[a]
900 " " 1000	17.6	5.0	39.9	14.4	23.2	100.0[a]
1000 " " 1100	17.5	4.9	38.9	15.1	23.7	100.0[a]
1100 " " 1200	16.6	4.7	37.7	14.9	26.1	100.0
1200 or over	17.4	5.0	36.5	15.7	25.4	100.0
All incomes	18.1	5.7	43.1	13.0	20.1	100.0

[a] The error of .1 in the sum of the items in the row is due to the use of only one decimal. For discussion of this point see Chapter XV, page 412.

3. Finally, the percentages of Table 81 are presented in graphic form in Figure 61. The bars are of equal length (100 per cent) and each is divided into component parts. The food and rent items are emphasized by placing them at the ends of the bars. A glance at the graph sug-

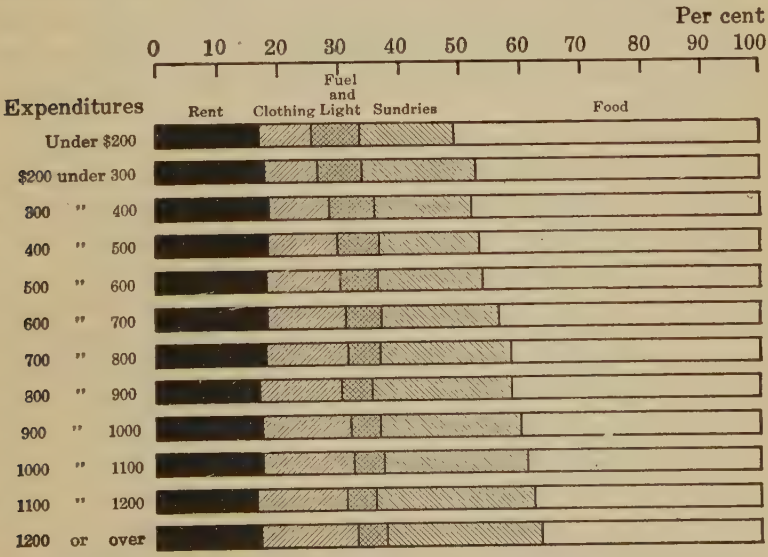


FIG. 61. PROPORTIONATE AVERAGE FAMILY EXPENDITURES FOR SPECIFIED PURPOSES, BY CLASSIFIED INCOMES, UNITED STATES, 1902

(Data from Table 81.)

gests that the proportion spent for rent remains about the same as the income increases, and that the food item decreases in relative importance.

The reader should compare the situation in 1902 as presented in Table 81 and Figure 61 with the situation in 1918–19 as shown in Table 82 and Figure 69, page 430.

USES AND ADVANTAGES OF GRAPHIC REPRESENTATION

The graph is a device to visualize and to throw into relief significant data and their relations.

1. *A graph relieves the mind of burdensome details* by keeping them in the background and by substituting points, lines and surfaces. *In most cases it makes vivid what is already known to the investigator himself.* To emphasize this aspect we say that the graph is a "representation." Great caution must be exercised by the maker of graphs to keep the picture true to the facts. On this account graphic methods deserve careful study. Even without deliberate intention it is easy to misrepresent the actual situation, or to distort the relations of data.

If the aim of a diagram is to prove something, it is a dangerous device, but as a means of portrayal it is an invaluable aid. To give the reader a wrong impression of the facts by an ignorant or careless use of a curve or a bar diagram is likely to produce the same effect as an error in the original data. Since a special appeal is made to the eye, wrong impressions received from diagrams are more vivid and more difficult to correct than when gained from tables.

2. A graphic device may be used as an *instrument of discovery* similar to the microscope. Sometimes factors and relationships in the problem which have escaped the observation of the investigator are thus revealed. It is difficult for the mind to compare simultaneously the details of several series of data. The graph enables the scientific worker to observe variations and movements within the same series and to relate these to other series (see Chapter XIII). In frequency distributions and in time series of daily, weekly, or monthly data the diagram preserves details which are concealed by averages and other statistical measures discussed in the chapters of Part II.

3. Graphic devices are useful in *popularizing the results of an investigation.* It is often desirable that the facts presented should arouse the interest of those not immediately concerned with the inquiry. The ordinary reader is averse to tables of figures because their meaning is usually not easily understood; but simple graphic devices, by emphasizing

specific phases of a problem, capture his attention. This initial interest often leads to the desire for more information and eventually to the making of intelligent public opinion. It is not enough, for example, to marshal the facts which indicate the causes of industrial accidents and the specific processes in different industries which are especially dangerous. The chief objective is to prevent accidents. To accomplish this purpose public interest must be aroused to the point of supporting legislation and enforcing specific standards of safety.

The busy executive welcomes the graphic representation of outstanding facts and relations in his organization, accompanied perhaps by the original data in tabular form for reference and by a minimum of explanatory notes. He is thereby enabled to observe trends and to make the comparisons needed for his decisions without too great sacrifice of time. In a time series it is easy to add new data to the diagram month by month and to compare current facts with past experience.

The responsibility for *planning* diagrams is especially important, since the purpose in many cases is to make statistical data interesting to those who have neither the critical ability nor the information needed to guard themselves against wrong impressions.

ILLUSTRATIONS OF THE FAULTY CONSTRUCTION OF GRAPHS

The faulty construction of diagrams leads to unwarranted inferences. Even though the student may never prepare graphs it is quite likely that he will use and interpret those constructed by others. A knowledge of sound principles of construction and cautions concerning the practices to be avoided are necessary in the training of every student in statistical methods.

1. **Omission of a zero base line.** In Figure 62 the vertical scale begins at 100 instead of at 0, thus cutting off from each bar a distance representing 100 points in the infant death-rate. This method of construction gives the impression of a more rapid decline in the rate than actually took place. The rate in 1911 was about one half that in 1890, but the height of the bar representing 1911 is only about one sixth that representing 1890. In fact the infant death-rate of New York City for 1923 (66) would be represented far below the 100 level on this diagram. If all the vertical bars had been drawn from zero as a base line, they would have shown by their relative heights the rates in correct proportion to each other. *In diagrams drawn on the natural scale*[1] *the zero base line should appear whenever practicable.* (See rules 3 and 4, pp. 439–40.)

[1] By natural scale we mean a scale based upon absolute numbers in contrast to one based upon ratios or relative numbers.

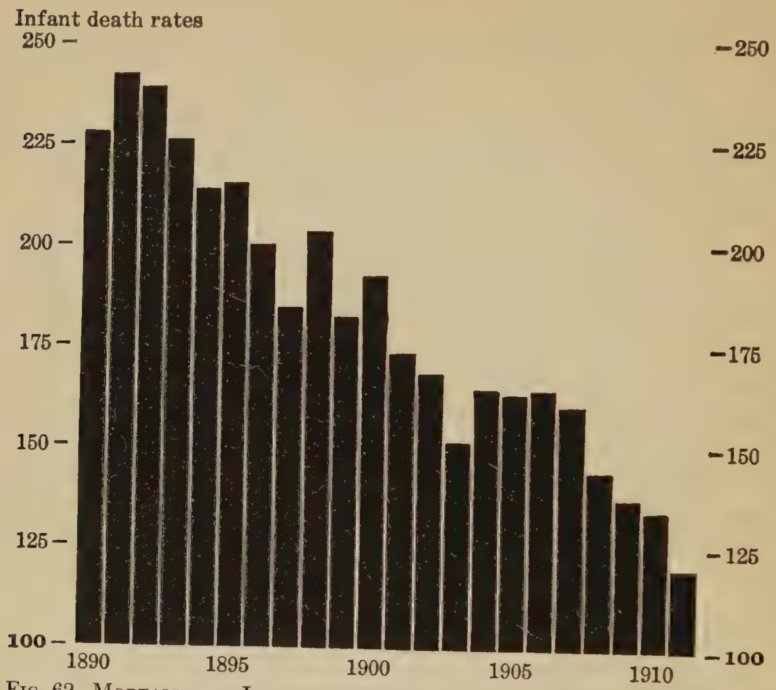


FIG. 62. MORTALITY OF INFANTS UNDER ONE YEAR OF AGE PER 1000 LIVING AT THAT AGE, NEW YORK CITY, 1890–1911

(Diagram has no zero base line. A distance should be added to each bar equal to 100 units on the vertical scale. From *Transactions of the Fifteenth International Congress on Hygiene and Demography*, Washington, 1912, vol. 6, p. 192.)

2. **The comparison of rectangular or bar diagrams.** In Figure 63A the upper bar is not the same width as the other two. It is the *area* of each bar which represents the number of plants. But the first impression of relative numbers is gained from the lengths, and since the lower bar representing 34 failures is wider, it is not comparable in length with the upper which shows the 58 plants in which scientific management was successful. This method of construction minimizes the failures by shortening the lower bar and making it wider. It happens that the author of the article in which this diagram appears wishes to present a strong case in favor of scientific management.

Figure 63B presents the same facts by the use of *bars of the same width.* To make a correct comparison the observer need only note the relative lengths. Linear magnitudes are more easily and accurately compared than areas or solids. (See rule 2, p. 439.)

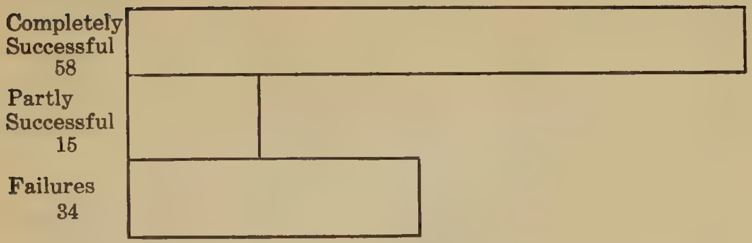


FIG. 63A. RECORD OF "SCIENTIFIC MANAGEMENT" IN 107 PLANTS

(From *System*, November, 1915, p. 456.)

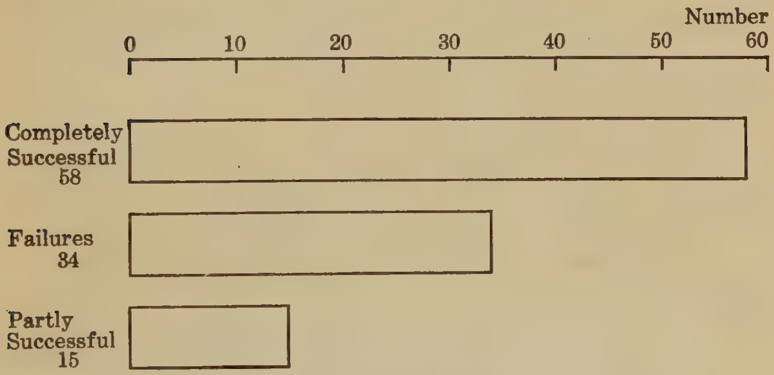


FIG. 63B. SAME DATA AS FOR FIGURE 63A

(Redrawn with rectangles all the same width, varying only in length, and with a measuring scale above the diagram.)

3. **Comparison of circles of varying size.** Circles of different sizes representing varying magnitudes are not compared as accurately as are bar diagrams, and should be avoided whenever the amount of variation is important. In Figure 64, page 424, the same data are portrayed by circles and by bars.

4. **Wrong impressions arising from the choice of scales.** The choice of the proper distance on the diagram to represent a given amount of variation in magnitude or a period of time is a matter for experimentation and judgment. It is very easy to minimize or to exaggerate variations by the selection of a particular scale. The graph is a *representation* of the facts and the picture should be a true one. The diagram should be sketched first in pencil, and if necessary more than one scale should be tried and the effect noted as in Figure 65 A and B, page 425.

5. **The use of pictorial representations.** In Figure 66A the heights of the men represent the increasing numbers securing employment year by

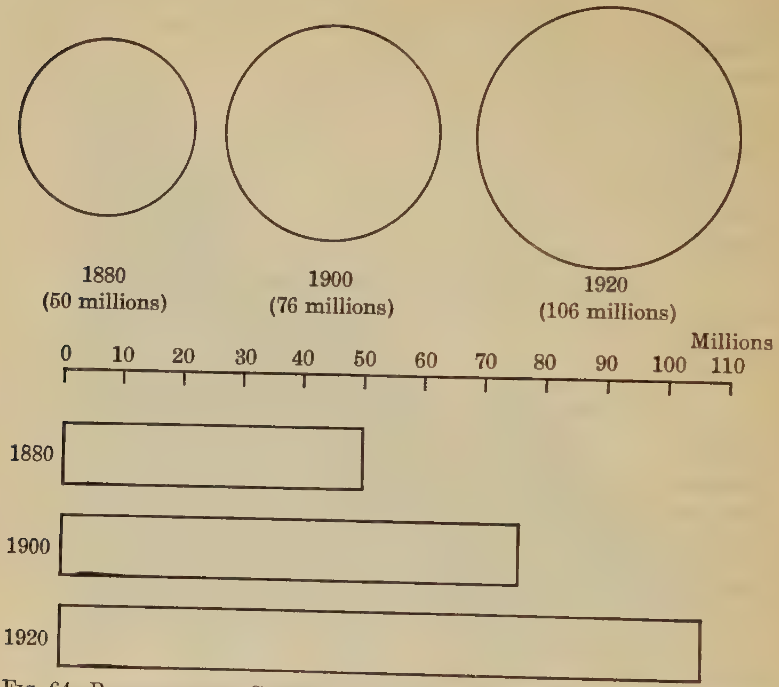


FIG. 64. POPULATION OF CONTINENTAL UNITED STATES, 1880, 1900, AND 1920
(Shown by circles of different areas and by bars varying in length.)

year. If the impression of relative numbers is gained from the pictures alone the situation for 1912–13, as compared with that for 1908–09, is greatly exaggerated. The mind compares not the relative heights of the pictorial representations, but their relative bulk. This amounts to comparing solids and is much less accurate than the comparison of linear magnitudes used in the bars of Figure 66B, page 426.

Pictorial devices should be avoided wherever exact comparisons are essential. They may be employed, however, merely for the purpose of ranking magnitudes in order of size, or to attract the attention of the reader to facts which are portrayed by other types of graphs.

6. Quantities entered within the field of the diagram. The numbers of deaths and the names of the months entered in Figure 67 just above the base line interfere with the interpretation of the diagram. The observer receives the impression that the base line is located above the numbers instead of below the names of the months. This minimizes greatly the apparent importance of the midsummer deaths relative to other seasons, reducing by more than one half the heights of the

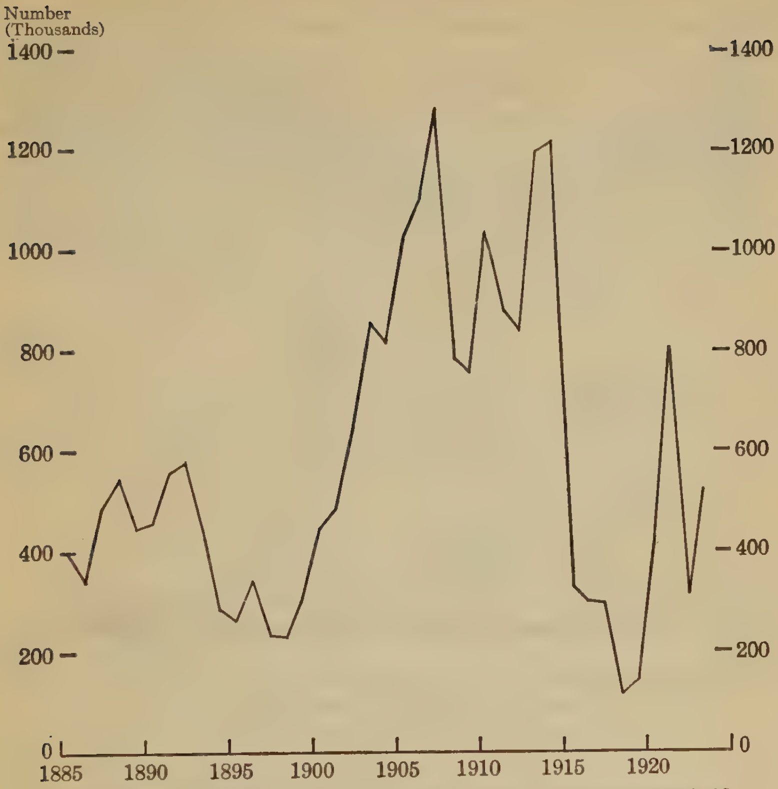


FIG. 65A. ANNUAL IMMIGRATION TO THE UNITED STATES, 1885–1923

(Data from *Statistical Abstract of the United States*, 1923, p. 75.)

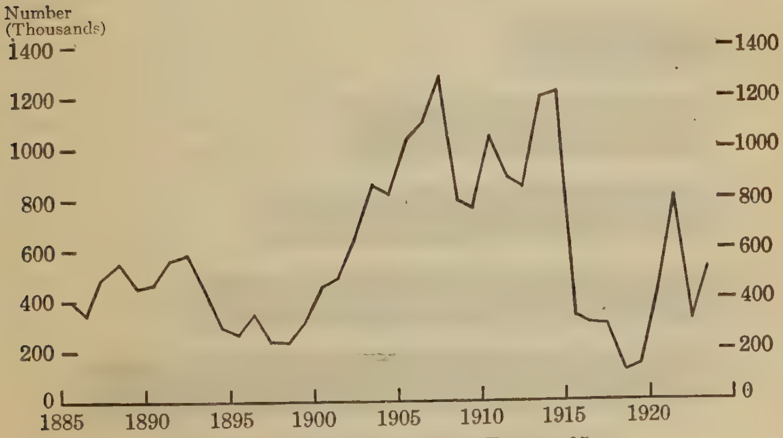


FIG. 65B. SAME DATA AS FOR FIGURE 65A

(Vertical scale one half that of A, horizontal scale the same.)

rectangular columns representing the mortality of July, August, and September.

It is desirable to have the detailed data which the diagram represents placed in a position convenient for reference, but *nothing should be per-*

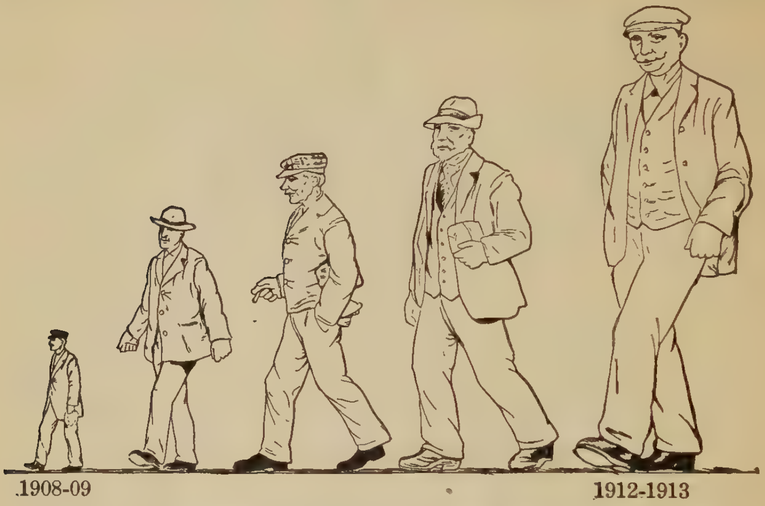


FIG. 66A. NUMBERS OF MEN FINDING WORK THROUGH THE DANISH UNSKILLED LABORERS' UNION, 1908–1912

(From *The Survey*, December 20, 1913, p. 337.)

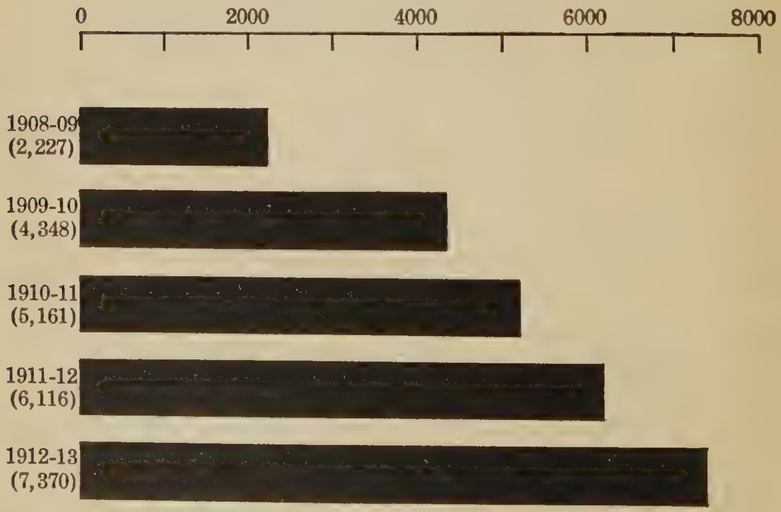


FIG. 66B. SAME DATA AS FOR FIGURE 66A

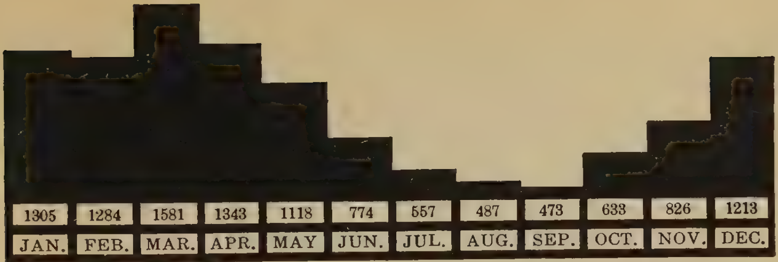


FIG. 67. NUMBER OF DEATHS FROM PNEUMONIA, BRONCHITIS, COLDS AND GRIPPE, NEW YORK CITY, BY MONTHS, 1915

(From *Weekly Bulletin* of the Department of Health, November 25, 1916.)

mitted to interfere with easy and accurate interpretation. Sometimes the data can be entered upon the diagram with advantage; in other cases they should be located in tabular form near the graph, on the same or an adjoining page (see rules 14 and 15, pp. 444–45).

THE FIELD UPON WHICH A GRAPH IS DRAWN — RECTANGULAR COÖRDINATES

Diagrams are representations of data by means of points, lines and surfaces the positions of which are located in space and described by reference to a *system of coördinates.* These coördinates are of different kinds depending upon the type of diagram, but for the most common graphs they are *rectangular*, as in Figure 68.

The position of the point *A* in Figure 68 is described with reference to the *X* and *Y* axes, as indicated by the dotted lines. *AC* is drawn parallel to the axis of ordinates. The distance *OC* or *BA* is known as the *abscissa* of the point *A*. Similarly *AB* is drawn parallel to the axis of abscissas. The distance *OB* or *CA* is known as the *ordinate* of the

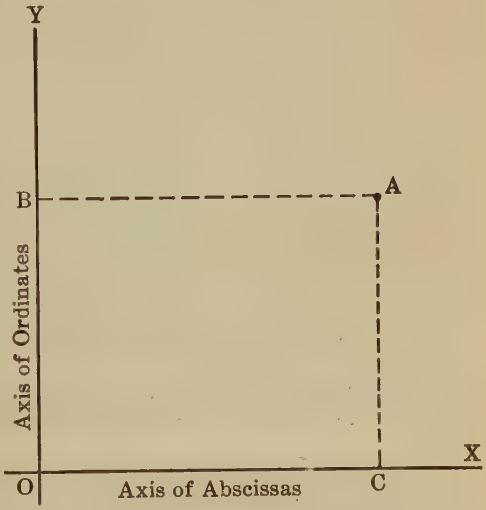


FIG. 68. THE FIELD UPON WHICH A DIAGRAM IS PLOTTED

(Illustrating Rectangular Coördinates.)

point A. In plotting a frequency polygon (Chapter V) the classes are usually laid off on the horizontal axis, or axis of abscissas, and the frequencies are plotted as ordinates on the vertical axis.

Deviations from the Y axis as origin toward the right and from the X axis as origin upward have *positive signs;* deviations to the left of OY and below OX as origins have *negative signs.*

TYPES OF DIAGRAMS

Before deciding to use a statistical diagram we should have a definite reason for preferring it to the tabular form of presentation, or for using it in addition to the table. Having settled this point the next step is to choose the type of graphic device best suited to the particular purpose and to the data. Is one kind more easily understood than another or more accurate in representing the facts? Is the same type of diagram equally appropriate for continuous data, as a frequency distribution of ages, and for representing magnitudes of independent categories, as the numbers of workers in different occupations? How should a time series be portrayed? What kind of graph is especially adapted to show geographic distributions? In representing increases and decreases are absolute amounts or relative values more important? The answers to these and other similar questions involve a classification of graphic devices. Limitation of space will not permit a comprehensive treatment in this book. Only the simpler and commonly used types are suggested, and the reader is referred for a full discussion to one or more of the treatises mentioned at the close of this chapter.

1. **Line diagrams.** The frequency polygon, the smooth frequency curve and the cumulative frequency diagram are of this type, illustrated in Chapters V–XI. The purpose of these diagrams is to represent the frequencies in magnitude classes of a characteristic, as income or age, which varies continuously on a scale of values from lowest to highest or *vice versa.* The histogram (Figure 6, p. 63) differs from the line diagrams in that it represents frequencies by rectangular areas, the width of each showing the size of the class-interval and the height indicating the class frequency. But when drawn as a *step diagram* (Figure 8, p. 69) the histogram is really similar to the other line diagrams.

Line diagrams are also especially adapted to the representation of time series (Chapter XIII). The continuity of the line emphasizes the sequence of the time units — weeks, months, or years.

Line diagrams are constructed with reference to rectangular coördinates (Figure 68). The horizontal and vertical scales are laid off on the respective axes X and Y, proceeding usually from left to right and from the bottom of the diagram toward the top. The scale numbers may be either *absolute*, as in the examples already cited in this chapter, or they may be the *logarithms* of the original numbers. The *logarithmic or ratio diagram* will be treated later as a special case.

2. **Area diagrams.** In a sense the frequency polygon and curve are area diagrams. It is the area above the horizontal axis bounded by the curve which represents the total frequency. With the exception of the histogram, however, frequencies are represented by ordinates and the relative number of cases in the different classes by the heights of these ordinates. The curve joining the tops of the ordinates indicates the varying number of cases in each class.

But there are magnitudes describing phenomena which, unlike frequency distributions of height and weight, *do not vary continuously*, for example, the populations of different cities or the average wage in different occupations. For these the curve or line diagram is not appropriate because it would suggest a continuity or sequence which does not exist in the data. Some other type should be used.

A. Bar diagrams. This is the simplest and most easily understood of all graphic devices, and is accurate in making comparisons. A variety of uses for this type is illustrated in the volume by Leonard P. Ayres, *The War with Germany*, in which a complicated story is told in clear graphic language. The bar is especially useful in representing data which do *not* vary continuously and which may be arranged in more than one order, as alphabetical or geographical, for example, the number of motor vehicles registered in each State or the annual value of products of different industries. This type of diagram is used frequently with good effect to set forth the data in a time series, the width of the bar representing the unit of time and the length showing the magnitude of the phenomenon. *It emphasizes the separate units of time*, as the amount of a commodity produced or shipped during a given week or month, *rather than the flow of time*, which is indicated more clearly by a curve. It is especially useful in describing data recorded *at specific times* with considerable intervals between records, as the amount of employment on the first of each month or the population at each decennial census.

To represent component parts as proportions of a whole the bar diagram is the most effective graphic device. The length of the bar represents 100 per cent. It is subdivided into two or more parts to show the proportions of the whole. Two or more bars, each divided into component parts, can be compared accurately and easily by placing them one above the other and by having a scale at the top extending from 0 to 100 per

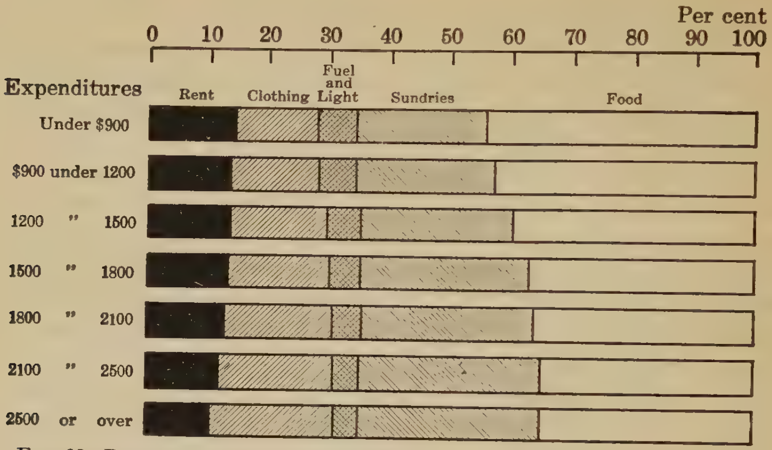


FIG. 69. PROPORTIONATE AVERAGE FAMILY EXPENDITURES FOR SPECIFIED PURPOSES, BY CLASSIFIED INCOMES, UNITED STATES, 1918–1919

(Data from Table 82.)

cent. Of course all bars are the same length since each represents 100 per cent, and they are more easily compared when placed horizontally on the page, one above the other, than in the vertical position. (See Figure 69 representing the data of Table 82.)

B. Circles — The sector diagram. Circles of different areas should not be used to compare varying frequencies or magnitudes. For this purpose the bar diagram is more effective. (Figure 64, p. 424.)

TABLE 82. PER CENT OF EXPENDITURES IN ONE YEAR FOR THE PRINCIPAL GROUPS OF ITEMS OF COST OF LIVING OF FAMILIES IN 92 INDUSTRIAL CENTERS, BY INCOME GROUPS, UNITED STATES, 1918–1919[a]

INCOME GROUP (1)	NUMBER OF FAMILIES (2)	RENT (3)	FUEL AND LIGHT (4)	FOOD (5)	CLOTHING (6)	SUNDRIES (7)	TOTAL (8)
Under $900	332	14.5	6.8	44.1	13.2	21.4	100.0
900 and under 1200	2,423	13.9	6.0	42.4	14.5	23.1	100.0
1200 " " 1500	3,959	13.8	5.6	39.6	15.9	25.0	100.0
1500 " " 1800	2,730	13.5	5.2	37.2	16.7	27.3	100.0
1800 " " 2100	1,594	13.2	5.0	35.7	17.5	28.5	100.0
2100 " " 2500	705	12.1	4.5	34.6	18.7	30.0	100.0
2500 and over	353	10.6	4.1	34.9	20.4	30.1	100.0
All incomes	12,096	13.0	5.2	38.2	16.6	26.4	100.0[b]

[a] Compiled from Bulletin 357, United States Bureau of Labor Statistics, Table 2, page 5. Furniture and Furnishing combined with Miscellaneous as Sundries.

[b] When the rows do not add exactly to 100.0, it is due to the use of only one decimal in the separate items.

The area of a circle, however, may be used to represent the whole, 100 per cent, as in the case of the bar, and may be subdivided into two or more sectors showing component parts. This is sometimes called a "pie" diagram. The circle conveys the impression of a whole very effectively, but the relation of the parts is not so clearly or accurately shown by the sectors as by the subdivided bar.

When several circles each divided into sectors are to be compared, they should be the same size, representing 100 per cent. In this case the sectors of the different circles representing specific parts of the whole are not as easily or accurately compared as are the subdivisions of the bars placed one above the other. Therefore, while the sector diagram is useful for variety in graphic presentation and for publicity work where

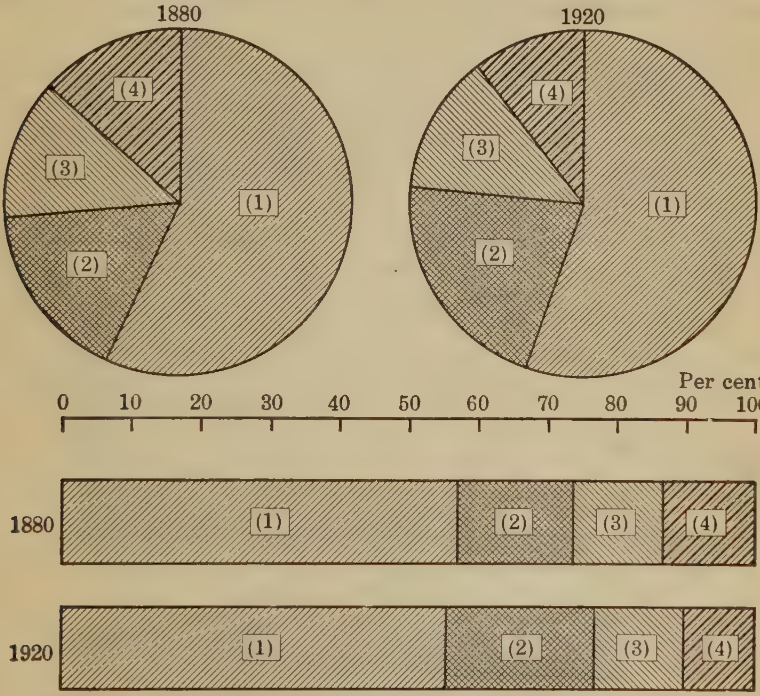


FIG. 70. NATIVITY OF POPULATION OF CONTINENTAL UNITED STATES 1880 AND 1920

Shown by sector and bar diagrams representing the same data as component parts of 100 per cent.

Key:	1880	1920
(1) Native white of native parentage..................	57.0	55.3 per cent
(2) Native white of foreign or mixed parentage.........	16.5	21.4 " "
(3) Foreign born white	13.1	13.0 " "
(4) Negro and other colored	13.4	10.3 " "
Total	100.0	100.0 " "

exact comparisons are not essential, *it should not be generally used to show component parts.* The bar diagram is to be preferred as illustrated in Figure 70.

3. **Maps.** Statistical maps are useful to represent geographic location or differences. The two kinds in most general use are the *pin or spot map* and the *shaded map.*

The shaded map is illustrated by Figure 2, page 49. The *geographic* differences in the distribution of the foreign population of the United States are shown by the shading. A key accompanies the diagram which makes clear to the reader the meaning of the different degrees of shading.[1]

The *geographic location* of a phenomenon or a frequency may be represented by a dot or other symbol, as in Figure 71. After the United States entered the World War 200,000 workmen were occupied continuously in army construction projects, which included piers and warehouses, plants for making explosives, repair shops, power plants, roads and housing for the troops. Their distribution in every State is shown in Figure 71. Each dot represents a separate construction project. Density of the dots on a map of this kind gives an impression of relative distribution over the area.

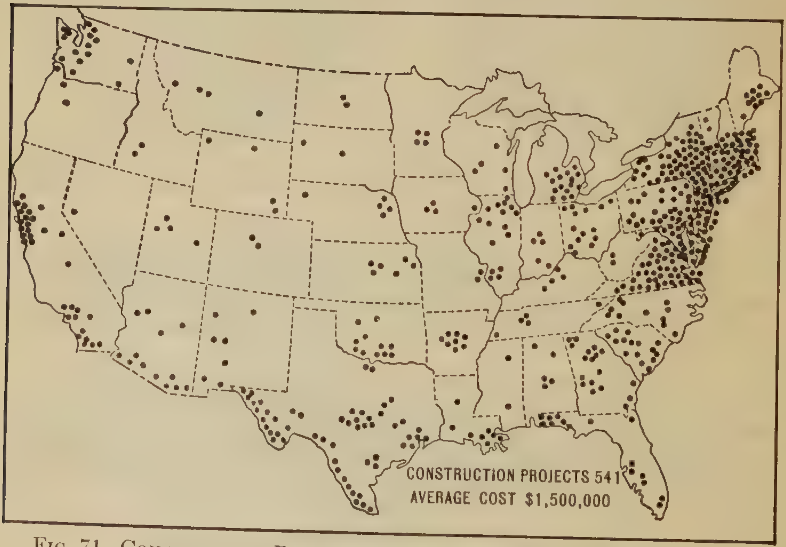


Fig. 71. Construction Projects of the Army in the United States during the World War

(From *The War with Germany*, Leonard P. Ayres, p. 57.)

[1] An excellent discussion of map making by Professor W. Z. Ripley is found in the *Quarterly Publication of the American Statistical Association*, September, 1899. The student should read this article.

A *pin map* is one especially constructed to permit the insertion of a pin representing each event or number of events at the proper location on the map. The heads of the pins are so designed that they distinguish by colors or other symbols different kinds of events, for example a street accident by passenger motor vehicle or by motor truck, cases of different contagious diseases under supervision, location of salesmen of rival firms in the same territory. The density of pins of the same kind gives an impression of the relative numbers in different locations.

4. The ratio diagram. Two magnitudes of like kind may be compared by noting their *difference.* The ordinary line diagram with scales based upon absolute amounts represents and compares differences. The same quantities may be compared by computing the *ratio* of the one to the other. *The purpose of a ratio diagram is to exhibit and to compare ratios.*

Often ratios of change or difference are more important than the absolute amounts. This is especially true of a time series. For example, the addition of a million immigrants to the population of the United States in 1790 would have been a 25 per cent increase, but the same number added to the present population would be less than one per cent. The two situations are very different from the point of view of growth and assimilation, yet the absolute increases are the same. Let us suppose that the sales of a small business grow during a single year from $10,000 to $20,000, an increase of 100 per cent. The same amount added to the business when it has reached $1,000,000 in size is a gain of only 1 per cent. During the period 1916 to 1920 both money wages and prices of commodities increased each year. Were *real wages* increasing or decreasing? The answer depends upon the *rate of increase* of money wages as compared with the *rate of increase* of prices.

To show relative variations between magnitudes in a time series a diagram with the *vertical* scale based upon absolute quantities is inadequate because equal spaces represent equal absolute amounts of difference. For this purpose it is necessary to represent equal ratios between magnitudes by equal vertical distances on the scale. Figures 72, 73, 74, 75, 76 and 77, illustrate the two methods of presenting data.

Equal vertical spacing in Figure 72 represents equal absolute amounts of difference anywhere on the diagram. Since the amount of increase (5000) is constant for each decade the lines are straight, and since the difference in the sizes of the two cities remain constant the lines are parallel. The distance of any point on either line above the zero base indicates the size of the city at that time but the *slope of the line between two points does not represent the rate of growth.* The slope of each line is the same throughout the period but the rate of growth is not constant. The one city starts at 5000 and doubles during the first decade (100 per cent) but increases only $14\frac{2}{7}$ per cent during the last decade. The other city, starting at 15,000, increases $33\frac{1}{3}$ per cent during the first decade and only $11\frac{1}{9}$ per cent during the last decade.

The fact that the lines are parallel does not mean that the rates of growth are the same for a particular decade. For example, the one city increases 100 per cent from 1850 to 1860 while the other increases only $33\frac{1}{3}$ per cent.

It is evident that in this type of diagram the slopes of two or more curves should not be compared, because they do not represent rates of increase or decrease. It is the *levels* of points on the curves which are comparable, and not the slopes. Yet the characteristics of curves which usually attract the interest of the observer are their comparative direction, their steepness or flatness, and whether or not they are parallel.

If the vertical spacing of Figure 72 were laid off to represent equal ratios of difference between successive magnitudes by equal spaces *the zero base line would cease to have any significance* and the horizontal ruling at 15,000 would be lowered in position until the distance between 10,000 and 15,000 (50 per cent increase) would be one half of the distance between 5000 and 10,000 (100 per cent increase). The distance between 10,000 and 20,000 would be the same as the distance between 5000 and 10,000, or between 20,000 and 40,000 (each being an increase of 100 per cent). When this principle is applied in constructing the vertical scale *equal vertical spaces anywhere on the diagram represent equal ratios of difference between magnitudes, the horizontal scale remaining as before.*

This is called a *ratio* vertical scale, a *semi-logarithmic* scale or an *arith-log* scale because the spacing represents ratios in the vertical direction and absolute amounts in the horizontal direction. Paper ruled in this manner may be purchased from concerns handling supplies for graphic work.

The data of Figure 72 are shown in Figure 73 on the ratio scale. The lines are no longer straight or parallel. Ratios of difference between successive magnitudes are depicted in each curve. The *slope* of the curve between any two points represents *rate of increase,* the same slope indicating the same rate anywhere on the diagram. When the rate declines the curve becomes concave on the lower side. *The slopes of two or more curves drawn on the ratio scale are comparable, representing comparative rates of increase or decrease.*

Figure 74 represents on absolute scales the populations of two States and of the United States at each census from 1850 to 1920. The curves

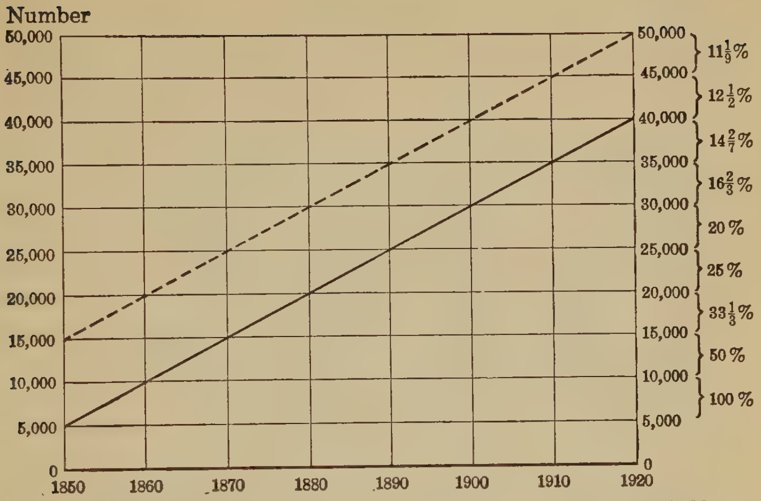


FIG. 72. GROWTH OF THE POPULATIONS OF TWO CITIES, 1850–1920

(Natural scale diagram. Each city increases 5000 per decade and this increase is represented by equal vertical distances on the diagram.)

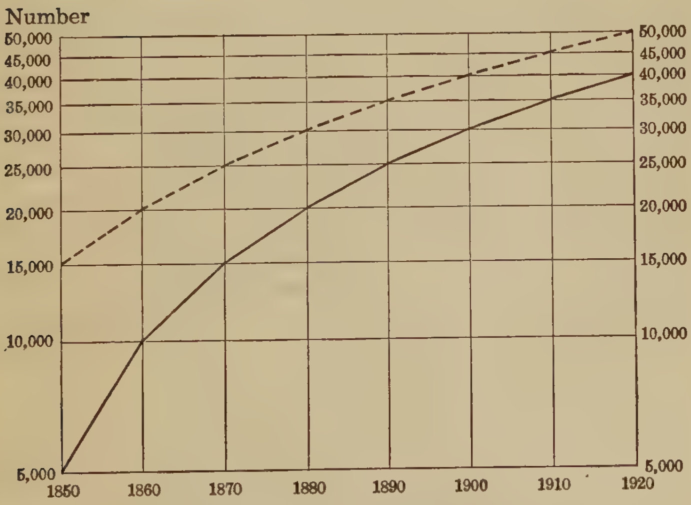


FIG. 73. COMPARISON OF THE RATES OF GROWTH OF POPULATIONS OF TWO CITIES, 1850–1920

(Same data as for Figure 72. Equal ratios of increase are represented by equal vertical distances. A ratio or semi-logarithmic scale diagram.)

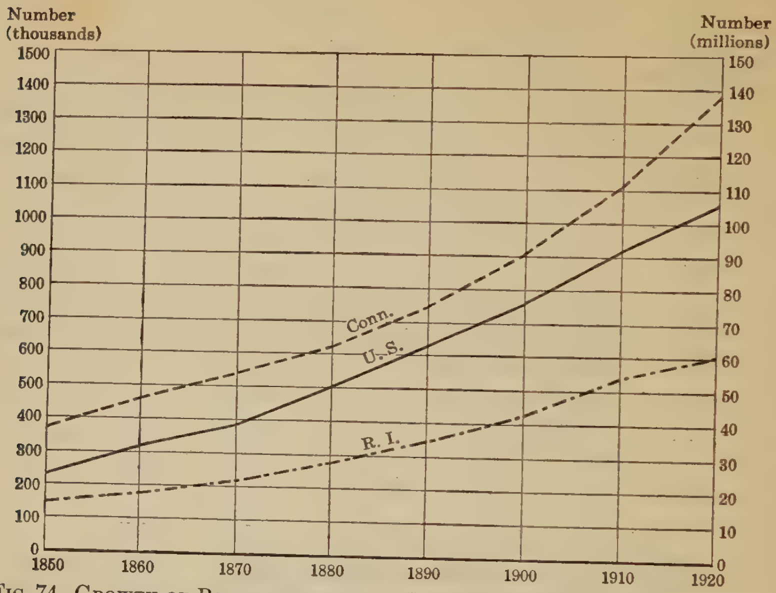


FIG. 74. GROWTH OF POPULATIONS OF THE UNITED STATES, OF RHODE ISLAND, AND OF CONNECTICUT, BY DECADES, 1850–1920

(Natural scale diagram. Two scales are used, the one on the left in thousands for the two states, and the other on the right in millions for the United States.)

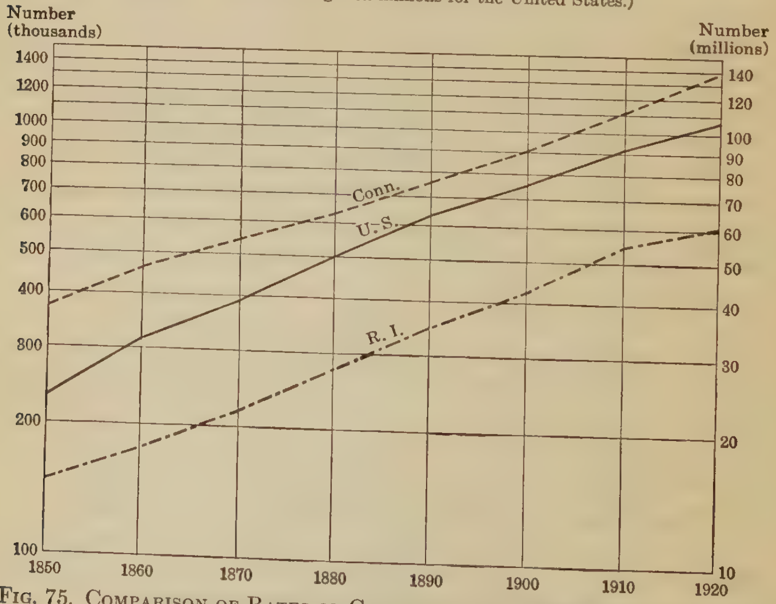


FIG. 75. COMPARISON OF RATES OF GROWTH OF THE UNITED STATES, RHODE ISLAND, AND CONNECTICUT, 1850–1920

(Same data as for Figure 74. A ratio or semi-logarithmic scale diagram.)

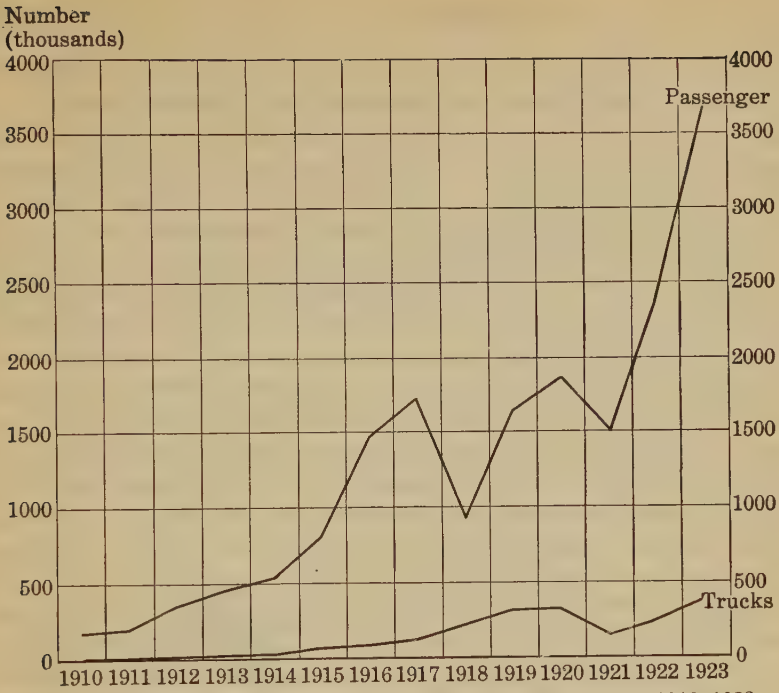


FIG. 76. NUMBER OF AMERICAN MOTOR VEHICLES PRODUCED, 1910–1923

(Natural scale diagram. Figures on production include the output of all plants in the United States and their branch factories in Canada. See *Commerce Reports*, United States Department of Commerce, May 19, 1924, p. 423.)

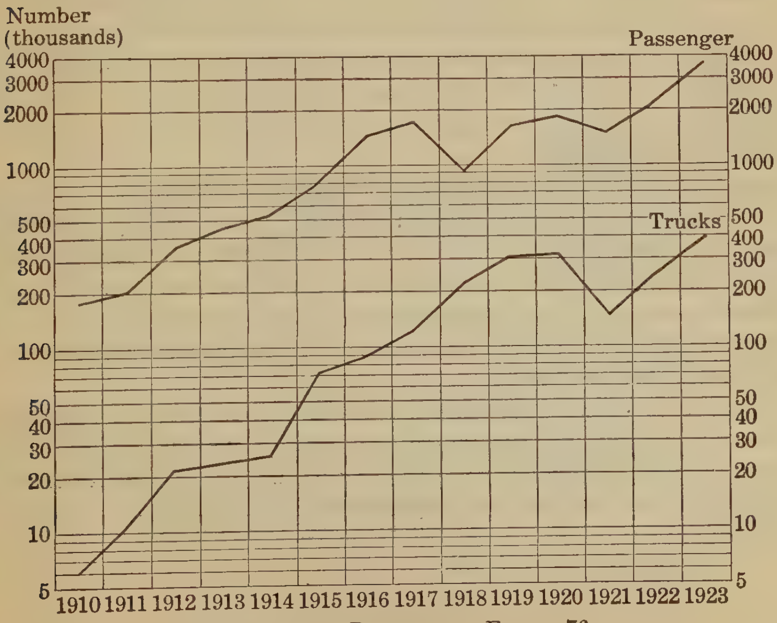


FIG. 77. SAME DATA AS FOR FIGURE 76

(Ratio or semi-logarithmic scale.)

for Rhode Island and Connecticut are plotted from the scale on the left, that for the United States from the scale on the right of the diagram. In order to present the large numbers for the entire country on the same graph a different scale is used. This diagram has little if any value and is likely to deceive the reader. It would be better to present the data in tabular form. Points on the State curves are comparable, but these are not comparable with points on the curve for the United States. *The slopes of the curves are not comparable since they do not represent rates of growth.*

Figure 75 shows the same data on the ratio scale. *The distance above the base line has no significance in this diagram.* Therefore, the large numbers for the United States offer no difficulty in plotting. Ciphers may be omitted or all the numbers may be reduced by a common divisor without affecting the ratios between them. *In this manner the curves may be brought close together for comparison.* It is the *slopes* that are comparable and not the *levels* of points above a base line. The curves reveal the relative rates of increase.

Figures 76 and 77 represent the growth in the production of passenger motor vehicles and motor trucks. The data are the same for both diagrams but the curves appear quite different when drawn on the ratio scale. The comparatively small numbers of motor trucks are easily represented in Figure 77 in relation to the larger numbers of passenger cars. The two curves of Figure 77 are comparable from the point of view of the *rates of increase or decrease* in the production of the two kinds of vehicles, regardless of whether the absolute numbers are large or small. The differences in per cents between successive magnitudes in each series are portrayed by the vertical distances on the diagram. A given per cent of increase or of decrease is always represented by the same vertical distance on the diagram.

In a ratio diagram if the curve is ascending and straight, or descending and straight, the *rate* of increase or decrease in the magnitudes is constant; if the curve bends upward the *rate* of growth is increasing, and if it bends downward the *rate* is decreasing. The *comparative steepness* of different portions of the same curve indicates *comparative rates* of increase or decrease. Parallel curves on the ratio diagram indicate equal rates of increase or decrease.

The use of the ratio scale is not confined to time series. Sometimes this scale is used in the horizontal direction as well as the vertical. For a more complete discussion of ratio diagrams and their applications the references at the close of this chapter should be consulted.

STANDARDS FOR GRAPHIC PRESENTATION[1]

1. The general arrangement of a diagram should proceed from left to right.

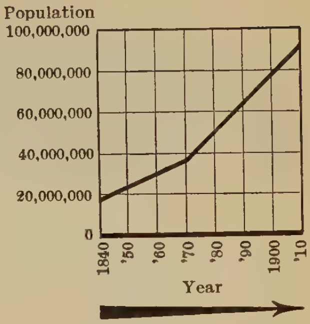


FIG. 1

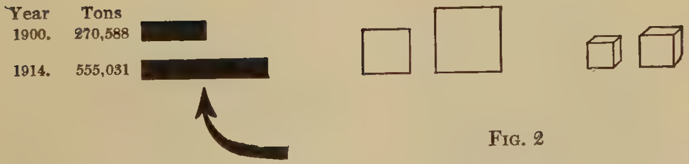


FIG. 2

2. Where possible represent quantities by linear magnitudes as areas or volumes are more likely to be misinterpreted.

3. For a curve the vertical scale, whenever practicable, should be so selected that the zero line will appear on the diagram.

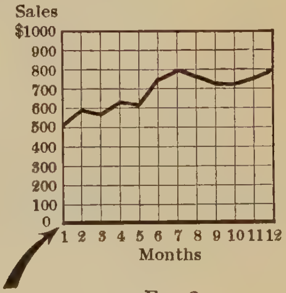


FIG. 3

[1] At the invitation of the American Society of Mechanical Engineers, and under the leadership of Willard C. Brinton, representatives of fifteen scientific societies and two government bureaus formed a Joint Committee on Standards for Graphic Presentation. In 1915, this committee published a report which is reprinted by permission in this volume. It sets forth "the more generally applicable principles of elementary graphic presentation." Reprinted from the *Quarterly Publication of the American Statistical Association*, December, 1915.

4. If the zero line of the vertical scale will not normally appear on the curve diagram, the zero line should be shown by the use of a horizontal break in the diagram.

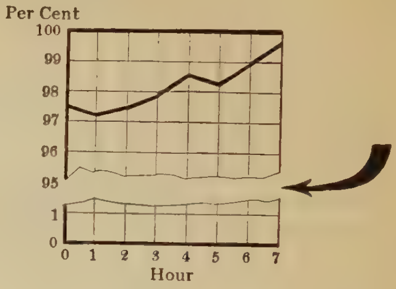


FIG. 4

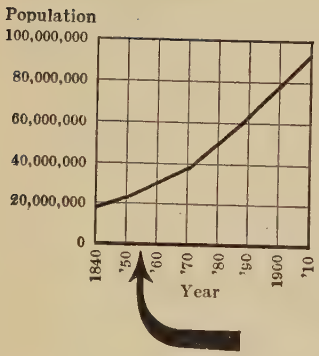


FIG. 5A

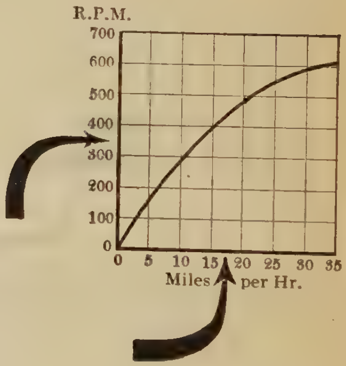


FIG. 5B

5. The zero lines of the scales for a curve should be sharply distinguished from the other coördinate lines.

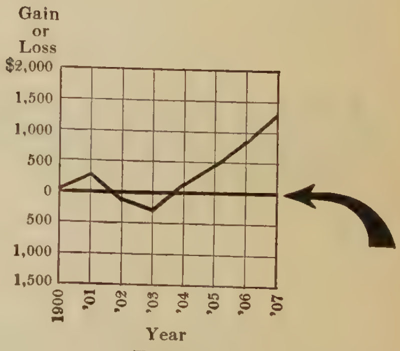


FIG. 5C

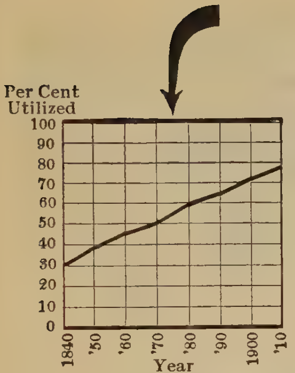


FIG. 6A

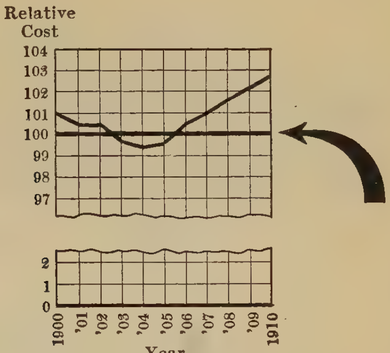


FIG. 6B

6. For curves having a scale representing percentages, it is usually desirable to emphasize in some distinctive way the 100 per cent line or other line used as a basis of comparison.

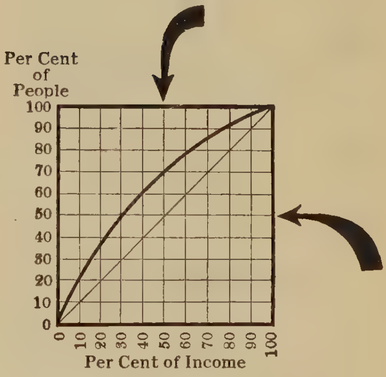


FIG. 6C

7. When the scale of a diagram refers to dates, and the period represented is not a complete unit, it is better not to emphasize the first and last ordinates, since such a diagram does not represent the beginning or end of time.

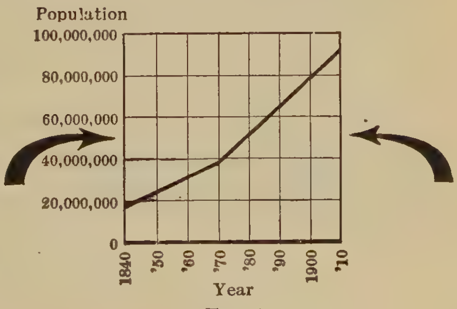


FIG. 7

8. When curves are drawn on logarithmic coördinates, the limiting lines of the diagram should each be at some power of ten on the logarithmic scales.

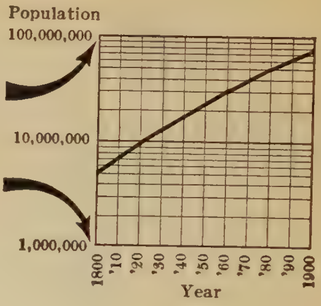


FIG. 8

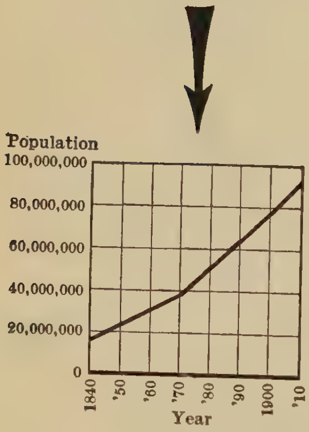


FIG. 9A

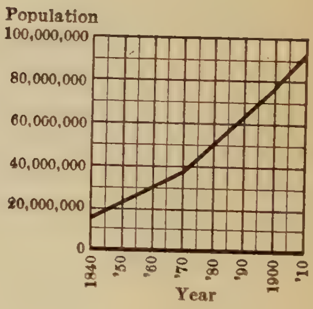


FIG. 9B

9. It is advisable not to show any more coördinate lines than necessary to guide the eye in reading the diagram.

10. The curve lines of a diagram should be sharply distinguished from the ruling.

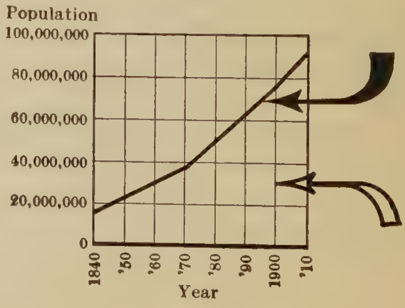


FIG. 10

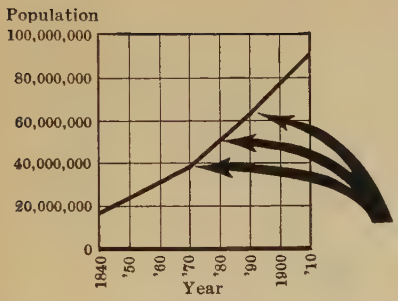


FIG. 11A

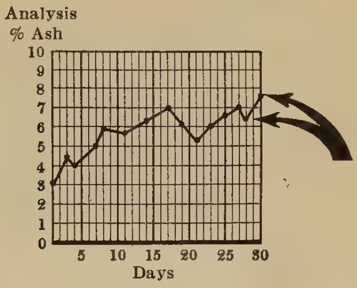


FIG. 11B

11. In curves representing a series of observations, it is advisable, whenever possible, to indicate clearly on the diagram all the points representing the separate observations.

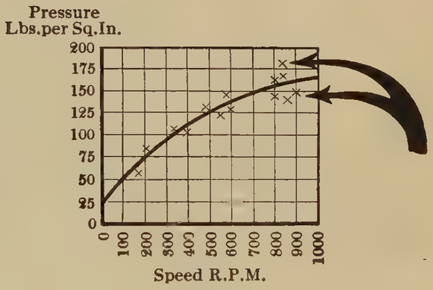


FIG. 11C

12. The horizontal scale for curves should usually read from left to right and the vertical scale from bottom to top.

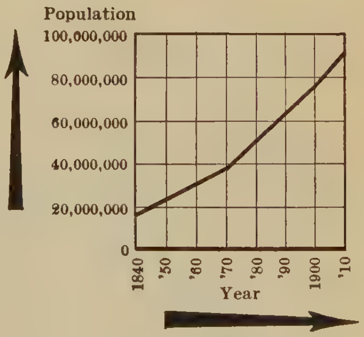


FIG. 12

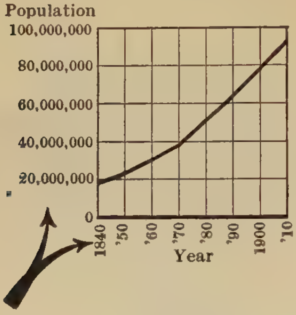


FIG. 13A

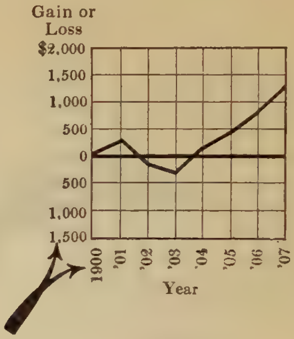


FIG. 13B

13. Figures for the scales of a diagram should be placed at the left and at the bottom or along the respective axes.

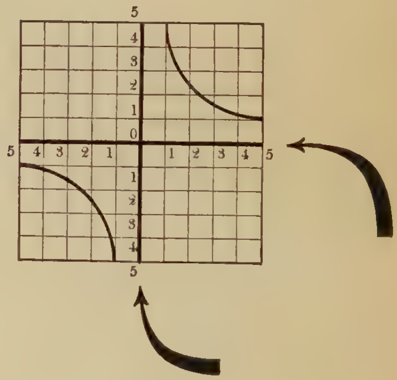


FIG. 13C

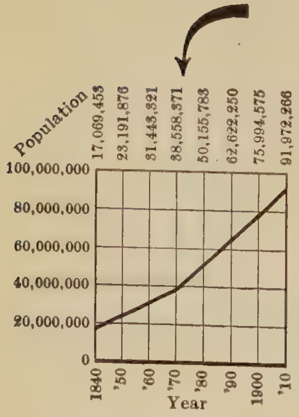


FIG. 14A

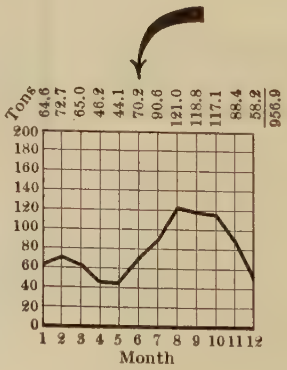


FIG. 14B

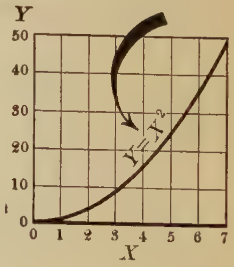
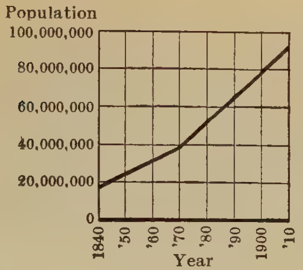


FIG. 14C

14. It is often desirable to include in the diagram the numerical data or formulæ represented.

15. If numerical data are not included in the diagram it is desirable to give the data in tabular form accompanying the diagram.



Year	Population
1840	17,069,453
1850	23,191,876
1860	31,443,321
1870	38,558,371
1880	50,155,783
1890	62,622,250
1900	75,994,575
1910	91,972,266

FIG. 15

16. All lettering and all figures on a diagram should be placed so as to be easily read from the base as the bottom, or from the right-hand edge of the diagram as the bottom.

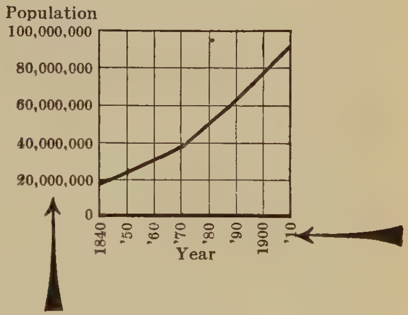
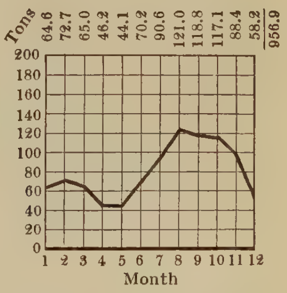


FIG. 16

17. The title of a diagram should be made as clear and complete as possible. Sub-titles or descriptions should be added if necessary to insure clearness.



ALUMINUM CASTINGS OUTPUT OF PLANT NO. 2, BY MONTHS, 1914

Output is given in short tons. Sales of scrap aluminum are not included

FIG. 17

READINGS

Karsten, Karl G., *Charts and Graphs.* (General treatise.)

Brinton, W. C., *Graphic Methods.* (General treatise.)

Jerome, Harry, *Statistical Method,* chaps. 4, 5, and 6. (Useful classifications of graphic devices, pp. 70 and 71, and a discussion of the Ratio Chart in chapter 6.)

Mills, F. C., *Statistical Methods as Applied to Economics and Business,* chap. 2. (Describes the straight line and other forms of curves including logarithmic.)

Pearl, Raymond, *Medical Biometry and Statistics,* chap. 6.

Whipple, G. C., *Vital Statistics,* 2d ed., chaps. 3 and 6. (In chapter 6 the problem of predicting future population is treated and many population curves for states and cities are presented.)

Bowley, A. L., *Elements of Statistics,* 4th ed., part I, chap. 7.

REFERENCES

Fisher, Irving, "The Ratio Chart," *Quarterly Publication of the American Statistical Association,* June, 1917. (An excellent treatment.)

Field, James A., "Some Advantages of the Logarithmic Scale in Statistical Diagrams," *Journal of Political Economy,* October, 1917. (The student is urged to read this and Professor Fisher's article.)

Bivins, P. A., "The Ratio Chart and Its Application," series of articles in *Industrial Management,* July, August, September, and October, 1921. (An excellent bibliography in the October issue.)

Ayres, L. P., *The War With Germany,* Government Printing Office, Washington, 1919. (The complex story of the War told in simple graphic language. Excellent examples of good graphic forms.)

Ripley, W. Z., "Notes on Map Making and Graphic Representation," *Quarterly Publication of the American Statistical Association,* September, 1899. (An excellent discussion of map shading and other details of graphic construction.)

Rorty, M. C., "Making Statistics Talk," series of articles in *Industrial Management,* December, 1920, and January and February, 1921.

Burnet, A. R., Series of articles in *Management Engineering,* August, September, November, and December, 1921. (Describes method of scale selection.)

Haskell, A. C., *How to Make and Use Graphic Charts.*

—— ——, *Graphic Charts in Business.*

Marshall, W. C., *Graphical Methods.*

Lipka, Joseph, *Graphical and Mechanical Computation.*

Publications of the Bureau of the Census, Washington. (Especially the *Statistical Atlas* of the Twelfth and of the Thirteenth Census.)

[Details as to publisher and date of publication of the readings and references given at the close of each chapter are found in the alphabetical list in Appendix A.]

APPENDIX A

ALPHABETICAL LIST OF REFERENCES

Bibliographical Note. No attempt is made to furnish a complete bibliography in this text. Selected readings and references have been given at the close of chapters. These are arranged in Appendix A in alphabetical order, each with the name of the publisher and the date of publication. The student will find excellent bibliographies in the following texts:

(1) Yule, *An Introduction to the Theory of Statistics*, 6th edition, 1922, at the close of each chapter and in the Supplements, pp. 387–92 (especially useful for references to original papers and discussions).

(2) *Handbook of Mathematical Statistics* (H. L. Rietz, Editor), pp. 195–208.

(3) Kelley, *Statistical Method*, Appendix B.

American Telegraph and Telephone Company, *Statistical Analysis and Projection of Time Series*, Statistical Bulletin No. 4, Statistical Method Series, Office of the Chief Statistician, New York, 1922.

An Introduction to Reflective Thinking, Columbia Associates in Philosophy, Houghton Mifflin, Boston, 1923.

Ayres, Leonard P., *The War With Germany*, Government Printing Office, Washington, 1919.

Bailey, W. B., and Cummings, John, *Statistics*, McClurg, Chicago, 1917.

Berridge, W. A., Winslow, E. A., and Flinn, R. A., *Purchasing Power of the Consumer — A Statistical Index*, A. W. Shaw Company, Chicago, 1925.

Bertillon, J., *Cours élémentaire de statistique*, Société d'éditions scientifiques, Paris, 1895.

Bivins, P. A., "The Ratio Chart and its Application," *Industrial Management*, July, August, September, and October, 1921.

Blakeman, J., "On Tests for Linearity of Regression in Frequency Distributions," *Biometrika*, Volume IV, 1905.

Bowley, A. L., *Elements of Statistics*, Fourth Edition, P. S. King and Son, London, 1920. (Charles Scribner's Sons, New York.)

—— *An Elementary Manual of Statistics*, Macdonald and Evans, London, 1910.

—— *The Measurement of Social Phenomena*, P. S. King and Son, London, 1915.

Bowley, A. L., and Burnett-Hurst, A. R., *Livelihood and Poverty*, G. Bell and Sons, London, 1915.

Brinton, W. C., *Graphic Methods*, Engineering Magazine Co., New York, 1914.

Brunt, David, *The Combination of Observations*, Cambridge University Press, 1917. (Putnam, New York.)

Burdge, H. G., *Our Boys, A Study of 245,000 Employed Boys of the State of New York*, Military Training Commission, Bureau of Vocational Training, Albany, 1921.

Burnet, A. R., Series of articles on graphic presentation in *Management Engineering*, August, September, November, December, 1921.

Carr, Elma B., "Cost of Living Statistics of the United States Bureau of Labor

Statistics and the National Industrial Conference Board," *Journal of the American Statistical Association*, December, 1924.

Carver, H. C., "Frequency Curves," in the *Handbook of Mathematical Statistics* (H. L. Rietz, Editor), chapter 7.

Chapin, F. S., *Field Work and Social Research*, The Century Co., New York, 1920.

Charlier, C. V. L., *Vorlesungen über die Grundzüge der Mathematischen Statistik*, Lund, 1920.

Coolidge, J. L., *An Introduction to Mathematical Probability*, Oxford University Press, London, 1925.

Crum, W. L., "Progressive Variation in Seasonality," *Journal of the American Statistical Association*, March, 1925.

Davies, G. R., *Introduction to Economic Statistics*, The Century Co., New York, 1922.

Day, E. E., "Classification of Statistical Series," *Quarterly Publication of the American Statistical Association*, December, 1919.

—— "Standardization of the Construction of Statistical Tables," *Quarterly Publication of the American Statistical Association*, March, 1920.

—— "An Index of the Physical Volume of Production," *Review of Economic Statistics*, September, 1920–January, 1921, Harvard University.

Douglas, Paul H., and Lamberson, Frances, "The Movement of Real Wages 1890–1918," *American Economic Review*, September, 1921.

Drachsler, Julius, *Intermarriage in New York City*, Studies in History, Economics and Public Law, Columbia University, Volume 94, No. 2, 1921.

Durand, E. Dana, "Tabulation by Mechanical Means," *Transactions of the Fifteenth International Congress on Hygiene and Demography*, 1912, Volume 6, pp. 83–90, Washington, 1913.

Edgeworth, F. Y., "Methods of Statistics," *Jubilee Volume of the Statistical Society of London* (now Royal Statistical Society), London, 1885.

—— "Probability," in Eleventh Edition of *The Encyclopædia Britannica*.

Elderton, W. P., and Ethel M., *Primer of Statistics*, Adam and Charles Black, London, 1910. (Macmillan, New York.)

Elderton, W. P., *Frequency Curves and Correlation*, C. and E. Layton, London, 1906.

Falk, I. S., *The Principles of Vital Statistics*, W. B. Saunders Co., Philadelphia, 1923.

Falkner, Helen D., "The Measurement of Seasonal Variation," *Journal of the American Statistical Association*, June, 1924.

Field, James A., "Some Advantages of the Logarithmic Scale in Statistical Diagrams," *Journal of Political Economy*, October, 1917.

Fisher, Irving, "The Ratio Chart," *Quarterly Publication of the American Statistical Association*, June, 1917.

—— *The Making of Index Numbers*, Houghton Mifflin, Boston, 1922.

—— "The Best Form of Index Number," *Quarterly Publication of the American Statistical Association*, March, 1921.

Flux, A. W., "The Measurement of Price Changes," *Journal of the Royal Statistical Society*, March, 1921.

Giddings, F. H., *The Scientific Study of Human Society*, University of North Carolina Press, Chapel Hill, 1924.

Giddings, F. H., "The Service of Statistics to Sociology," *Quarterly Publication of the American Statistical Association*, March, 1914.

—— "The Measurement of Social Forces," *The Journal of Social Forces*, November, 1922.

Giffen, Robert, *Statistics*, Macmillan, London, 1913.

Griffin, F. L., *An Introduction to Mathematical Analysis*, Houghton Mifflin, Boston, 1921.

Hall, L. W., "Seasonal Variation as a Relative of Secular Trend," *Journal of the American Statistical Association*, June, 1924.

Hansen, A. H., "The Buying Power of Labor During the War," *Journal of the American Statistical Association*, March, 1922.

Hart, Hornell, "Science and Sociology," *The American Journal of Sociology*, November, 1921.

Hart, W. L., "The Method of Monthly Means for Determination of a Seasonal Variation," *Journal of the American Statistical Association*, September, 1922.

Haskell, A. C., *How to Make and Use Graphic Charts*, Codex Book Co., New York, 1919.

—— *Graphic Charts in Business*, Codex Book Co., New York, 1922.

Hexter, M. B., *Social Consequences of Business Cycles*, Houghton Mifflin, Boston, 1925.

Hilton, John, "Enquiry by Sample: An Experiment and Its Results" (Discussion by Yule, Bowley, Edgeworth and Greenwood), *Journal of the Royal Statistical Society*, July, 1924.

Hoffman, F. L., *Insurance Science and Economics*, The Spectator Company, New York, 1911.

Hollerith, H., "An Electric Tabulating System," *School of Mines Quarterly*, Columbia University, April, 1889. (Beginnings of mechanical methods in census work.)

Huntington, E. V., "Curve Fitting by the Method of Least Squares and the Method of Moments," *Handbook of Mathematical Statistics* (H. L. Rietz, Editor), chapter 4.

Jerome, Harry, *Statistical Method*, Harper and Brothers, New York, 1924.

John, V., *Geschichte der Statistik*, Enke, Stuttgart, 1884.

Jones, Adam Leroy, *Logic Inductive and Deductive. An Introduction to Scientific Method*, Henry Holt, New York, 1909.

Jones, D. C., *A First Course in Statistics*, G. Bell and Sons, London, 1921.

Jordan, D. F., *Business Forecasting*, Prentice-Hall, New York, 1923.

Karsten, Karl G., *Charts and Graphs*, Prentice-Hall, New York, 1923.

Kelley, Truman L., *Statistical Method*, Macmillan, New York, 1923.

Keynes, J. M., *A Treatise on Probability*, Macmillan, London, 1921.

King, W. I., *Elements of Statistical Method*, Macmillan, New York, 1912.

—— "An Improved Method for Measuring the Seasonal Factor," *Journal of the American Statistical Association*, September, 1924.

Koren, John (Editor), *The History of Statistics. Their Development and Progress in Many Countries*, Macmillan, New York, 1918.

Lipka, Joseph, *Graphical and Mechanical Computation*, John Wiley and Sons, New York, 1918.

Manual of the International List of Causes of Death, Department of Commerce, Bureau of the Census, Washington, 1924.

Marshall, W. C., *Graphical Methods*, McGraw-Hill Book Co., New York, 1921.

Mayo-Smith, Richmond, *Statistics and Sociology*, Macmillan, New York, 1895.

—— *Statistics and Economics*, Macmillan, New York, 1899.

Meitzen, August, *History, Theory and Technique of Statistics*. Translated by Roland P. Falkner, Annals of the American Academy of Political and Social Science, Supplement, March, 1891.

Merriman, M., *A Textbook on the Method of Least Squares*, John Wiley and Sons, New York, 1910.

Merz, J. T., "On the Statistical View of Nature," *A History of European Thought in the Nineteenth Century*, Second unaltered edition, Volume II, chapter 12, Wm. Blackwood and Sons, London, 1912.

Methods of Procuring and Computing Statistical Information of the Bureau of Labor Statistics (wages, cost of living, accidents, prices and employment), Bulletin 326, United States Bureau of Labor Statistics, Washington, 1923.

Mills, Frederick C., *Statistical Methods as Applied to Economics and Business*, Henry Holt, New York, 1924.

—— "On Measurement in Economics" from *The Trend of Economics* (R. G. Tugwell, Editor), Alfred A. Knopf, New York, 1924.

—— "The Measurement of Correlation and the Problem of Estimation," *Journal of the American Statistical Association*, September, 1924.

Mitchell, H. H., and Grindley, H. S., *The Element of Uncertainty in the Interpretation of Feeding Experiments*, Bulletin 165, Agricultural Experiment Station, University of Illinois, July, 1913.

Mitchell, Wesley C., "Quantitative Analysis in Economic Theory," *American Economic Review*, March, 1925.

—— *Index Numbers of Wholesale Prices in the United States and Foreign Countries*, Bulletin 284, United States Bureau of Labor Statistics, Washington, 1921. (Revision of Bulletin 173, 1915.)

—— *Business Cycles*, California University Publications, Volume III, 1913.

—— *History of Prices During the War*, Price Bulletin No. 1, War Industries Board, Washington, 1919.

Moore, Henry L., *Laws of Wages*, Macmillan, New York, 1911.

—— *Economic Cycles: Their Law and Cause*, Macmillan, New York, 1914.

—— *Forecasting the Yield and the Price of Cotton*, Macmillan, New York, 1917.

—— *Generating Economic Cycles*, Macmillan, New York, 1923.

National Bureau of Economic Research, *Income in the United States* (Wesley C. Mitchell, Editor), 2 volumes, Harcourt Brace and Co., New York, 1921.

Newcomb, H. T., "The Development of Mechanical Methods of Statistical Tabulation," *Transactions of the Fifteenth International Congress on Hygiene and Demography*, 1912, Volume 6, pp. 73–83, Washington, 1913.

Newsholme, Sir Arthur, *The Elements of Vital Statistics*, New Edition, D. Appleton, New York, 1924.

Ogburn, W. F., and Thomas, Dorothy, "Influence of the Business Cycle on Certain Social Conditions," *Quarterly Publication of the American Statistical Association*, September, 1922.

Pearl, Raymond, *Medical Biometry and Statistics*, W. B. Saunders Co., Philadelphia, 1923.

Pearson, Karl, *The Grammar of Science*, Part I — Physical, Third Edition, Adam and Charles Black, London, 1911. (Macmillan, New York.)

Pearson, Karl, "On the General Theory of Skew Correlation and Non-linear Regression," *Drapers' Company Research Memoirs:* Biometric Series II, London, 1905.

—— *Tables for Statisticians and Biometricians*, Cambridge University Press, 1914.

Persons, Warren M., "Correlation of Time Series," *Journal of the American Statistical Association*, June, 1923. (*Handbook of Mathematical Statistics*, Chapter 10.)

—— *The Review of Economic Statistics*, Preliminary Volume I, Harvard University, 1919. (Methods of treating time series.)

—— "Construction of a Business Barometer," *American Economic Review*, December, 1916.

—— "The Variate Difference Correlation Method and Curve Fitting," *Quarterly Publication of the American Statistical Association*, June, 1917.

—— "Fisher's Formula for Index Numbers," *Review of Economic Statistics*, Preliminary Volume III, 1921, pp. 103–113, Harvard University.

Pollak Foundation for Economic Research, *The Problem of Business Forecasting* (Edited by Persons, Foster and Hettinger). Papers presented at the eighty-fifth annual meeting of the American Statistical Association, December, 1923, published by Houghton Mifflin, Boston, 1924.

Riegel, Robert, *Elements of Business Statistics*, D. Appleton, New York, 1924.

Rietz, H. L. (Editor), *Handbook of Mathematical Statistics*, Houghton Mifflin, Boston, 1924.

Rietz, H. L., and Crathorne, A. R., "Simple Correlation," in the *Handbook of Mathematical Statistics*, chapter 8.

Ripley, W. Z., "Notes on Map Making and Graphic Representation," *Quarterly Publication of the American Statistical Association*, September, 1899.

Ross, Frank A., *School Attendance in the United States in 1920*, Census Monographs V, Bureau of the Census, Washington, 1924. (Correlation Method, Appendix A.)

Rossiter, W. S., *A Century of Population Growth*, Bureau of the Census, Washington, 1909.

Rugg, H. O., *Statistical Methods Applied to Education*, Houghton Mifflin, Boston, 1917.

Secrist, Horace, *An Introduction to Statistical Methods*, Macmillan, New York, 1917.

—— *Readings and Problems in Statistical Methods*, Macmillan, New York, 1920.

Standardization of Industrial Accident Statistics, Bulletin 276, United States Bureau of Labor Statistics, Washington, 1920.

Thorndike, E. L., *An Introduction to the Theory of Mental and Social Measurements*, Second Edition, Teachers College, Columbia University, 1913.

—— *Individuality*, Houghton Mifflin, Boston, 1911.

Walsh, C. M., *The Measurement of General Exchange Value*, Macmillan, New York, 1901.

—— *The Problem of Estimation*, P. S. King, London, 1921.

Watkins, G. P., "Theory of Statistical Tabulation," *Quarterly Publication of the American Statistical Association*, December, 1915.

West, Carl J., *Introduction to Mathematical Statistics*, R. G. Adams, Columbus, 1918.

—— "Value to Economics of Formal Statistical Methods," *Quarterly Publication of American Statistical Association*, September, 1915.

Weld, L. D., *Theory of Errors and Least Squares*, Macmillan, New York, 1916.

Westergaard, Harald, "Scope and Method of Statistics," *Quarterly Publication of the American Statistical Association*, September, 1916.

—— *Die Lehre von der Mortalität und Morbilität*, Fischer, Jena, 1901.

Whipple, G. C., *Vital Statistics*, Second Edition, John Wiley and Sons, New York, 1923.

Whitaker, E. T., and Robinson, G., *The Calculus of Observations*, Blackie and Son, London, 1924.

Willcox, Walter F., "The Need of Social Statistics as an Aid to the Courts," *Quarterly Publication of the American Statistical Association*, March, 1913.

—— "Development of the American Census Office Since 1890," *Political Science Quarterly*, September, 1914.

—— "The Statistical Work of the United States Government," *Quarterly Publication of the American Statistical Association*, March, 1915.

Young, A. A., "Index Numbers," in the *Handbook of Mathematical Statistics*, (H. L. Rietz, Editor), Chapter 12.

—— "The Measurement of Changes of the General Price Level," *The Quarterly Journal of Economics*, Volume 35 (1921), pp. 557–573.

—— "Fisher's The Making of Index Numbers," *The Quarterly Journal of Economics*, Volume 37, February, 1923.

Yule, G. U., *An Introduction to the Theory of Statistics*, Sixth Edition, Charles Griffin and Company, London, 1922. (J. B. Lippincott, Philadelphia.)

—— "On the Time Correlation Problem, with Especial Reference to the Variate Difference Correlation Method," *Journal of the Royal Statistical Society*, July, 1921.

Zizek, Franz, *Statistical Averages*. Translated by Warren M. Persons, Henry Holt, New York, 1913.

APPENDIX B

SUMMARY OF SYMBOLS, EQUATIONS, AND FORMULÆ

(With a note on aids in computation)

Aids in Computation. Calculating machines are an invaluable aid in statistical work but their cost frequently places them beyond the reach of the individual student. For ordinary simple computing work the slide-rule is very useful and sufficiently accurate. When greater exactness in multiplying and dividing is required and when complicated formulæ are employed logarithms are almost essential. Five-figure or seven-figure tables will usually give the necessary degree of accuracy. The extended multiplication tables cited below will be found very useful and inexpensive. Barlow's *Tables of Squares, Cubes, and Reciprocals* are invaluable for the elementary student.

(1) Barlow's *Tables of Squares, Cubes, Square-roots, Cube-roots, and Reciprocals of all Integer Numbers up to 10,000*, E. and F. N. Spon, London (Spon and Chamberlain, New York).

(2) Peters, J., *Neue Rechentafeln für Multiplikation und Division*, G. Reimer, Berlin. (Products up to 100 × 10,000. Introduction in English.)

(3) Crelle, A. L., *Rechentafeln*, G. Reimer, Berlin. (Products up to 1,000 × 1,000.)

(4) Pearson, Karl, *Tables for Statisticians and Biometricians*, Cambridge University Press, 1914.

(5) Glover, J. W., *Tables of Applied Mathematics in Finance, Insurance, Statistics*, George Wahr, Ann Arbor, 1923.

(6) Miner, J. R., *Tables of* $\sqrt{1 - r^2}$ *and* $1 - r^2$ *for Use in Partial Correlation and in Trigonometry*, The Johns Hopkins Press, Baltimore, 1922.

(7) Ross, F. A., "Formulæ for Facilitating Computations in Time Series Analyses," *Journal of the American Statistical Association*, March, 1925.

Symbols, Equations and Formulæ

Symbols are employed only to a limited extent in this text with the purpose of centering the attention upon processes and their meaning. Many teachers will prefer to expand the use of symbols in the practice work of their students. The most important are grouped below under appropriate general topics treated in the text. Sometimes the same symbol is used with more than one meaning but in these cases the context should make clear the particular use. Following the list of symbols under each general topic the chief formulæ will be given.

1. Symbols used in the description of frequency distributions

X: a variable magnitude in a series of observations.

Y: a variable magnitude in a second series, employed, for example, in correlation where two series are compared and related.

N: total number of cases in a series of observations or in a frequency distribution.

f: the number of cases in a single class of a frequency distribution.

m: the value of the mid-point of a class.

Σ: a general symbol meaning *the sum of*. (ΣX means the sum of N observations in a series of X's.)

M: the arithmetic average or mean.

W: weights employed in combining or averaging a series of values.

$G.A.$: the value of an assumed mean. Used in the short method of computing the mean, standard deviation, and the coefficient of correlation.

x: deviation from the average. The difference between any magnitude in a series of X's and the average of the entire series (for example, $X - M = x$), or between the mid-value of a class interval and the average of the entire series ($m - M = x$). In a series of Y's the deviations are designated by y.

d: deviation plus or minus from the *guessed average* ($G.A.$) in *intervals* or *steps*. Used in computing the mean, standard deviation, and coefficient of correlation by the short method. The subscripts d_X and d_Y denote particular series.

c: a correction factor. The difference between the true mean and the assumed mean ($M - G.A. = c$). Used in computing the mean, standard deviation, and coefficient of correlation by the short method. The subscripts c_X and c_Y denote particular series. (On page 139 c is used in a different meaning, to denote the size of the class interval in locating the mode within the modal class.)

Q_1: the first or lower quartile.

Q_3: the third or upper quartile.

l: the lower limit of the modal class.

f_1: the frequency in the class just below the modal class.

f_2: the frequency in the class just above the modal class.

Q: the semi-interquartile range — the quartile deviation.

$A.D.$: the mean deviation, measured from median, mean, or mode.

σ: the standard deviation — the root mean square deviation about the mean.

V: the coefficient of variation, a measure of relative dispersion based upon the standard deviation.

$V_{A.D.}$: a measure of relative dispersion based upon the mean deviation.

V_Q: a measure of relative dispersion based upon the quartile deviation.

Formulæ and Processes

Simple Mean $= \frac{\Sigma X}{N}$. Data ungrouped and need not be in order of magnitude.

Weighted Mean $= \frac{\Sigma XW}{\Sigma W}$.

Simple Mean $= \frac{\Sigma mf}{N}$. Data grouped and computation by long method.

Mean (short method) $= G.A. +$ (c times size of interval), in which

$$c = \frac{\Sigma fd}{N} \text{ in intervals or steps, and}$$

$$\text{Mean} = G.A. + \left(\frac{\Sigma fd}{N} \text{ times size of interval}\right).$$ Care must be taken to make the summation algebraic and to preserve signs.

Median:

(1) In ungrouped data arranged in order of magnitude the median is the value of the middle item if the number is odd, and if the number of items is even it is approximately the average value of the two middle items.

better: middle 2 observations

(2) In grouped data we count one half the number of items in the entire distribution $\left(\frac{N}{2}\right)$ and locate the class in which the median falls. For the method of locating the median value within the class the reader is referred to pages 109–11 and to Figure 5.

Geometric Mean $= \sqrt[N]{X_1 \text{ times } X_2 \text{ times } \ldots \text{ times } X_N}$, and

$$\log \text{Geometric Mean} = \frac{\Sigma \log X}{N}.$$

The arithmetic mean of the logarithms of the separate numbers gives the logarithm of their geometric mean. The corresponding natural number is the geometric mean.

Mode:

(1) Locate the class of greatest frequency. Test this location in the distribution by regrouping and shifting (pages 131–32 of the text). Locate the most probable value within the *modal class*. The crudest method is to take the mid-value of the modal class.

(2) Weight the location within the modal class by the formula,

$$\text{Mode} = l + \frac{f_2}{f_2 + f_1}\ c,$$

which gives influence to the frequencies of classes adjoining the modal class.

(3) Base location upon the relation of the mean and median to the mode, using the formula,

$$\text{Mode} = \text{Mean} - 3\ (\text{Mean} - \text{Median}).$$

(4) May use method of smoothing the frequencies by a moving average (page 133 of text) in order to locate the maximum ordinate of the frequency curve.

(5) Mathematical methods of curve fitting to locate the maximum ordinate are outside the scope of this text. Refined methods are usually applicable only to continuous data.

Absolute Measures of Dispersion:

$$Q = \frac{Q_3 - Q_1}{2}.$$

$$A.D. = \frac{\Sigma fx}{N}.$$ The signs of x, the deviations, are disregarded.

$$\sigma = \sqrt{\frac{\Sigma x^2}{N}}.$$ Data ungrouped and computation by long method.

$$\sigma = \sqrt{\frac{\Sigma fx^2}{N-1}}.$$ Data grouped and computation by long method.

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - c^2}.$$ Data grouped and short method used.

The c is the same as used for the mean. This formula gives σ in intervals and the result is reduced to the original units of the problem by multiplying by the size of the interval.

Measures of Relative Dispersion:

$$V = \frac{\sigma \text{ times } 100}{\text{Mean}}$$

$$V_{A.D.} = \frac{A.D. \text{ times } 100}{\text{Median, Mean, Mode}}$$

$$V_Q = \frac{\frac{Q_3 - Q_1}{2} \text{ times } 100}{\frac{Q_3 + Q_1}{2}} = \frac{Q_3 - Q_1 \text{ times } 100}{Q_3 + Q_1}$$

$$\text{Skewness} = \frac{\text{Mean} - \text{Mode}}{\text{Standard Deviation}}, \text{ or,}$$

$$\text{Skewness} = \frac{3\ (\text{Mean} - \text{Median})}{\text{Standard Deviation}}, \text{ or,}$$

$$\text{Skewness} = \frac{(Q_3 - \text{Median}) - (\text{Median} - Q_1)}{\frac{Q_3 - Q_1}{2}} = \frac{Q_1 + Q_3 - 2\ \text{Median}}{\frac{Q_3 - Q_1}{2}}$$

2. Symbols employed in describing the distribution of errors

x: any distance laid off on the horizontal axis (X) from the mean as zero, plus or minus, in units of standard deviation $\left(\frac{x}{\sigma}\right)$.

y_o: the maximum ordinate erected at the mean as zero origin, in the symmetrical bell-shaped frequency curve.

y: any other ordinate in the bell-shaped curve erected at distances plus or minus from the mean laid off in units of σ.

e: base of the Napierian logarithms (2.71828).

π: ratio of the circumference of a circle to the diameter (3.1416).

σ_M: the standard error of the mean of a sample, due to the conditions of sampling.

$\sigma_{(standard\ deviation)}$: the standard error of the standard deviation of a sample.

σ_r: the standard error of the coefficient of correlation.

$\sigma_{(M_1 - M_2)}$: the standard error of the difference between two means.

$P.E.$: the probable error, due to the conditions of sampling (.6745 σ).

$P.E._M$: the probable error of the mean ($P.E._M = .6745\ \sigma_M$). Similarly, the symbol $P.E.$ with the corresponding subscript represents the probable error of the measure to which the subscript relates. In each case it is .6745 times the standard error.

±: a symbol indicating that the probable error is measured plus or minus from the statistical constant to which it relates.

Equations and Formulæ

$Y = mX + b$: the equation of a straight line.

$x^2 + y^2 = a^2$: the equation of a circle.

$y = \frac{N}{\sigma\sqrt{2\pi}} e^{\frac{-x^2}{2\sigma^2}}$: the equation of the bell-shaped symmetrical curve.

$y = y_o e^{\frac{-x^2}{2\sigma^2}}$: another form of the equation of the bell-shaped curve.

$y_o = \frac{N}{\sigma\sqrt{2\pi}} = \frac{N}{2.5066\ \sigma}$: the maximum ordinate of the bell-shaped curve, erected at the mean of the distribution.

$\sigma_M = \frac{\sigma_{\text{sample}}}{\sqrt{N}}$: the standard error of the mean of a sample.

$\sigma_{\text{(standard deviation)}} = \frac{\sigma_{\text{sample}}}{\sqrt{2\,N}}$: the standard error of the standard deviation of a sample.

$\sigma_{(M_1-M_2)} = \sqrt{(\sigma_{M_1})^2 + (\sigma_{M_2})^2}$: the standard error of a difference.

$P.E. = .6745\ \sigma$: the probable error is .6745 times the standard error.

$P.E._M = .6745\ \frac{\sigma_{\text{sample}}}{\sqrt{N}}$: the probable error of the mean of a sample.

$P.E._{\text{(standard deviation)}} = .6745\ \frac{\sigma_{\text{sample}}}{\sqrt{2N}}$: the probable error of the standard deviation of a sample.

$P.E._{(M_1-M_2)} = .6745\sqrt{(\sigma_{M_1})^2 + (\sigma_{M_2})^2}$: the probable error of a difference.

3. Symbols used in describing and measuring relationship

X: the value of a variable in a series of observations.

Y: the value of a second variable in a series of observations.

N: the total of related pairs.

$\overline{X}$: the mean of the entire X series.

$\overline{Y}$: the mean of the entire Y series.

x: deviation from the mean of the X series ($X - \overline{X} = x$).

y: deviation from the mean of the Y series ($Y - \overline{Y} = y$).

σ_X: the standard deviation of the X series.

σ_Y: the standard deviation of the Y series.

d_X: deviation in intervals from the G.A. of the X series, used in the short method.

d_Y: deviation in intervals from the $G.A.$ of the Y series, used in the short method.

c_X: correction in intervals for the $G.A.$ of the X series ($G.A. + c =$ Mean).

c_Y: correction in intervals for the $G.A.$ of the Y series.

σ_r: the standard error of the Pearsonian coefficient of correlation.

$P.E._{\cdot r}$: the probable error of the Pearsonian coefficient.

m_1: the slope of the straight line of average relationship of Y on X $\left(m_1 = r\,\frac{\sigma_Y}{\sigma_X}\right)$. The coefficient of regression of Y on X.

m_2: the slope of the straight line of average relationship of X on Y $\left(m_2 = r\frac{\sigma_X}{\sigma_Y}\right)$. The coefficient of regression of X on Y.

b: the constant in the general equation to a straight line ($Y = mX + b$), denoting the distance from zero of the point where the straight line cuts the Y axis.

S_Y: the measure of scatter, or the standard error of estimate about the line of average relationship Y on X.

S_X: the standard error of estimate about the line of average relationship X on Y.

η_{YX}: the correlation ratio of Y on X.

η_{XY}: the correlation ratio of X on Y.

$\overline{Y}_X$: the mean of any column of Y's corresponding to given values of X.

n_X: total frequencies in any column of Y's.

$\overline{X}_Y$: the mean of any row of X's corresponding to given values of Y.

n_Y: total frequencies in any row of X's.

$P.E._\eta$.: the probable error of the correlation ratio.

$\sigma_{(\eta^2 - r^2)}$: the standard error of the difference between η^2 and r². This is a test of the linearity of relationship.

ρ: Spearman's coefficient of correlation, based upon the squares of differences in rank.

D: the difference between the ranks of two related variables in Spearman's method of correlation by ranks.

Equations and Formulæ

$y = m_1x$: equation of the straight line of relationship Y on X, origin at the crossing of the means of the system of related values.

$x = m_2y$: equation of the straight line X on Y.

$m_1 = \frac{y}{x} = \left(r\frac{\sigma_Y}{\sigma_X}\right)$: the slope of the straight line Y on X.

$m_2 = \frac{x}{y} = \left(r\frac{\sigma_X}{\sigma_Y}\right)$: the slope of the straight line X on Y.

$y = r\frac{\sigma_Y}{\sigma_X}x$: equation of the straight line Y on X, in which x and y are deviations from the respective means of the series.

$x = r\frac{\sigma_X}{\sigma_Y}y$: equation of the straight line X on Y.

$Y - \bar{Y} = r\frac{\sigma_Y}{\sigma_X}(X - \bar{X})$: the equation to the straight line Y on X, in which X and Y are specific values in the respective series.

$X - \bar{X} = r\frac{\sigma_X}{\sigma_Y}(Y - \bar{Y})$: the equation to the straight line X on Y.

$r = \frac{\Sigma xy}{N\sigma_X\sigma_Y}$: the formula for the Pearsonian coefficient of correlation.

$r = \frac{\frac{\Sigma d_X d_Y}{N} - c_X c_Y}{\sigma_X \sigma_Y}$: short method formula for r.

$\sigma_r = \frac{1 - r^2}{\sqrt{N}}$: the standard error of the correlation coefficient.

$P.E._r = .6745\,\frac{1 - r^2}{\sqrt{N}}$: the probable error of the correlation coefficient.

$S_Y = \sigma_Y\sqrt{1 - r^2}$: measure of scatter about the straight line of relationship Y on X. Measure of the standard error of estimate.

$S_X = \sigma_X\sqrt{1 - r^2}$: measure of scatter about the straight line X on Y.

$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \text{ times } \Sigma y^2}}$: formula for computing r when the data are not tabulated in a correlation table, and when the regression equations are not desired.

$r = \sqrt{1 - \frac{S_Y^2}{\sigma_Y^2}}$: formula for r derived from scatter formula of Y on X.

$r = \sqrt{1 - \frac{S_X^2}{\sigma_X^2}}$: formula for r derived from scatter formula of X on Y.

$\eta_{YX} = \frac{\sqrt{\frac{\Sigma n_X(\bar{Y}_X - \bar{Y})^2}{N}}}{\sigma_Y} = \frac{\sigma \text{ of the means of the columns}}{\sigma \text{ of the entire } Y \text{ distribution}}$: the correlation ratio of Y on X.

$\eta_{XY} = \frac{\sqrt{\frac{\Sigma n_Y(\bar{X}_Y - \bar{X})^2}{N}}}{\sigma_X} = \frac{\sigma \text{ of the means of the rows}}{\sigma \text{ of the entire } X \text{ distribution}}$: the correlation ratio of X on Y.

$P.E._\eta = .6745\,\frac{1 - \eta^2}{\sqrt{N}}$: the probable error of the correlation ratio.

$\sigma_{(\eta^2 - r^2)} = 2\sqrt{\frac{\eta^2 - r^2}{N}}$: the abridged Blakeman formula for measuring the standard error of the difference between η^2 and r^2. Used in testing for non-linearity.

$\rho = 1 - \frac{6\Sigma D^2}{N\,(N^2 - 1)}$: Spearman's measure of correlation by ranks.

$r = 2 \sin\left(\frac{\pi}{6}\rho\right)$: Pearson's formula for correlation of grades or ranks. Used in transforming Spearman's ρ into r.

4. Symbols used in describing time series

X: a series of observations located at specific periods of time; or the units of time, used in correlating a variable with time in fitting a straight line to the data of a time series to describe the secular trend.

Y: a second series of observations located at specific periods of time.

x: deviations from the average of a series of X's; or from the values of the secular trend in order to eliminate the influence of the trend.

y: deviations from the average of a series of Y's; or from the values of the secular trend.

$\frac{x}{\sigma}$: deviations of a series of X's expressed in units of standard deviation.

$\frac{y}{\sigma}$: deviations of a series of Y's expressed in units of standard deviation.

c_X : correction used in correlation of a series of X's where the trend is described by a moving average and the deviations plus and minus do not balance. A similar correction (c_Y) is needed in a similar series of Y's.

m: the slope of the straight line describing secular trend.

Formulæ

$m = \frac{\Sigma xy}{\Sigma x^2}$: slope of the fitted straight line describing secular trend. Method of correlation with time which is the same as the method of least squares.

$r = \frac{\Sigma\left(\frac{x}{\sigma} \text{ times } \frac{y}{\sigma}\right)}{N}$: formula for correlating two time series when the deviations are already expressed in units of standard deviation.

Link relatives $= \frac{\text{Jan.}}{\text{Dec.}} \times 100; \frac{\text{Feb.}}{\text{Jan.}} \times 100$; etc. Used in obtaining an index of seasonal variation.

APPENDIX C[1]

Ordinates of the normal probability curve expressed as fractional parts of the mean ordinate y_o. Each ordinate is erected at a given distance from the mean. The height of the ordinate erected at the mean can be computed from,

$$y_o = \frac{N}{\sigma\sqrt{2\pi}} = \frac{N}{2.5066\,\sigma}$$

The corresponding height of any other ordinate can be read from the table by assigning the distance that the ordinate is from the mean (x). Distances on x are measured as fractional parts of σ. Thus the height of an ordinate at a distance from the mean of .7 σ will be .78270 y_o; the height of an ordinate at 2.15 σ from the mean will be .09914 y_o, etc.

x/σ	0	1	2	3	4	5	6	7	8	9
0.0	100000	99995	99980	99955	99920	99875	99820	99755	99685	99596
0.1	99501	99396	99283	99158	99025	98881	98728	98565	98393	98211
0.2	98020	97819	97609	97390	97161	96923	96676	96420	96156	95882
0.3	95600	95309	95010	94702	94387	94055	93723	93382	93024	92677
0.4	92312	91399	91558	91169	90774	90371	89961	89543	89119	88688
0.5	88250	87805	87353	86896	86432	85962	85488	85006	84519	84060
0.6	83527	83023	82514	82010	81481	80957	80429	79896	79359	78817
0.7	78270	77721	77167	76610	76048	75484	74916	74342	73769	73193
0.8	72615	72033	71448	70861	70272	69681	69087	68493	67896	67298
0.9	66689	66097	65494	64891	64287	63683	63077	62472	61865	61259
1.0	60653	60047	59440	58834	58228	57623	57017	56414	55810	55209
1.1	54607	54007	53409	52812	52214	51620	51027	50437	49848	49260
1.2	48675	48092	47511	46933	46357	45783	45212	44644	44078	43516
1.3	42956	42399	41845	41294	40747	40202	39661	39123	38569	38058
1.4	37531	37007	36487	35971	35459	34950	34445	33944	33447	32954
1.5	32465	31980	31500	31023	30550	30082	29618	29158	28702	28251
1.6	27804	27361	26923	26489	26059	25634	25213	24797	24385	23978
1.7	23575	23176	22782	22392	22008	21627	21251	20879	20511	20148
1.8	19790	19436	19086	18741	18400	18064	17732	17404	17081	16762
1.9	16448	16137	15831	15530	15232	14939	14650	14364	14083	13806
2.0	13534	13265	13000	12740	12483	12230	11981	11737	11496	11259
2.1	11025	10795	10570	10347	10129	09914	09702	09495	09290	09090
2.2	08892	08698	08507	08320	08136	07956	07778	07604	07433	07265
2.3	07100	06939	06780	06624	06471	06321	06174	06029	05888	05750
2.4	05614	05481	05350	05222	05096	04973	04852	04734	04618	04505
2.5	04394	04285	04179	04074	03972	03873	03775	03680	03586	03494
2.6	03405	03317	03232	03148	03066	02986	02908	02831	02757	02684
2.7	02612	02542	02474	02408	02343	02280	02218	02157	02098	02040
2.8	01984	01929	01876	01823	01772	01723	01674	01627	01581	01536
2.9	01492	01449	01408	01367	01328	01288	01252	01215	01179	01145
3.0	01111	00819	00598	00432	00309	00219	00153	00106	00073	00050
4.0	00034	00022	00015	00010	00006	00004	00003	00002	00001	00001
5.0	00000									

[1] From Rugg's *Statistical Methods Applied to Education*. Reprinted by arrangement with the publishers, Houghton Mifflin Company.

APPENDIX D[1]

Fractional parts of the total area (10,000) under the normal probability curve, corresponding to distances on the baseline between the mean and successive points of division laid off from the mean. Distances are measured in units of the standard deviation, σ. To illustrate, the table is read as follows: between the mean ordinate, y_o, and any ordinate erected at a distance from it of, say, $.8\,\sigma$ $\left(i.e., \frac{x}{\sigma} = .8\right)$, is included 28.81 per cent of the entire area.

x/σ	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	0000	0040	0080	0120	0159	0199	0239	0279	0319	0359
0.1	0398	0438	0478	0517	0557	0596	0636	0675	0714	0753
0.2	0793	0832	0871	0910	0948	0987	1026	1064	1103	1141
0.3	1179	1217	1255	1293	1331	1368	1406	1443	1480	1517
0.4	1554	1591	1628	1664	1700	1736	1772	1808	1844	1879
0.5	1915	1950	1985	2019	2054	2088	2123	2157	2190	2224
0.6	2257	2291	2324	2357	2389	2422	2454	2486	2518	2549
0.7	2580	2612	2642	2673	2704	2734	2764	2794	2823	2852
0.8	2881	2910	2939	2967	2995	3023	3051	3078	3106	3133
0.9	3159	3186	3212	3238	3264	3289	3315	3340	3365	3389
1.0	3413	3438	3461	3485	3508	3531	3554	3577	3599	3621
1.1	3643	3665	3686	3718	3729	3749	3770	3790	3810	3830
1.2	3849	3869	3888	3907	3925	3944	3962	3980	3997	4015
1.3	4032	4049	4066	4083	4099	4115	4131	4147	4162	4177
1.4	4192	4207	4222	4236	4251	4265	4279	4292	4306	4319
1.5	4332	4345	4357	4370	4382	4394	4406	4418	4430	4441
1.6	4452	4463	4474	4485	4495	4505	4515	4525	4535	4545
1.7	4554	4564	4573	4582	4591	4599	4608	4616	4625	4633
1.8	4641	4649	4656	4664	4671	4678	4686	4693	4699	4706
1.9	4713	4719	4726	4732	4738	4744	4750	4758	4762	4767
2.0	4773	4778	4783	4788	4793	4798	4803	4808	4812	4817
2.1	4821	4826	4830	4834	4838	4842	4846	4850	4854	4857
2.2	4861	4865	4868	4871	4875	4878	4881	4884	4887	4890
2.3	4893	4896	4898	4901	4904	4906	4909	4911	4913	4916
2.4	4918	4920	4922	4925	4927	4929	4931	4932	4934	4936
2.5	4938	4940	4941	4943	4945	4946	4948	4949	4951	4952
2.6	4953	4955	4956	4957	4959	4960	4961	4962	4963	4964
2.7	4965	4966	4967	4968	4969	4970	4971	4972	4973	4974
2.8	4974	4975	4976	4977	4977	4978	4979	4980	4980	4981
2.9	4981	4982	4983	4984	4984	4984	4985	4985	4986	4986
3.0	4986.5	4987	4987	4988	4988	4988	4989	4989	4989	4990
3.1	4990.3	4991	4991	4991	4992	4992	4992	4992	4993	4993
3.2	4993.129									
3.3	4995.166									
3.4	4996.631									
3.5	4997.674									
3.6	4998.409									
3.7	4998.922									
3.8	4999.277									
3.9	4999.519									
4.0	4999.683									
4.5	4999.966									
5.0	4999.997133									

[1] From Rugg's *Statistical Methods Applied to Education.*

APPENDIX E[1]

Values of r for corresponding values of ρ computed from the expressions,

$$\rho = 1 - \frac{6 \Sigma D^2}{N(N^2 - 1)}$$

$$r = 2 \sin \left(\frac{\pi}{6} \rho \right)$$

Values of r given in this table have been computed for various values of ρ ranging from .01 to 1.00.

ρ	r	ρ	r	ρ	r	ρ	r
.01	.0105	.26	.2714	.51	.5277	.76	.7750
.02	.0209	.27	.2818	.52	.5378	.77	.7847
.03	.0314	.28	.2922	.53	.5479	.78	.7943
.04	.0419	.29	.3025	.54	.5580	.79	.8039
.05	.0524	.30	.3129	.55	.5680	.80	.8135
.06	.0628	.31	.3232	.56	.5781	.81	.8230
.07	.0733	.32	.3335	.57	.5881	.82	.8325
.08	.0838	.33	.3439	.58	.5981	.83	.8421
.09	.0942	.34	.3542	.59	.6081	.84	.8516
.10	.1047	.35	.3645	.60	.6180	.85	.8610
.11	.1151	.36	.3748	.61	.6280	.86	.8705
.12	.1256	.37	.3850	.62	.6379	.87	.8799
.13	.1360	.38	.3935	.63	.6478	.88	.8893
.14	.1465	.39	.4056	.64	.6577	.89	.8986
.15	.1569	.40	.4158	.65	.6676	.90	.9080
.16	.1674	.41	.4261	.66	.6775	.91	.9173
.17	.1778	.42	.4363	.67	.6873	.92	.9269
.18	.1882	.43	.4465	.68	.6971	.93	.9359
.19	.1986	.44	.4567	.69	.7069	.94	.9451
.20	.2091	.45	.4669	.70	.7167	.95	.9543
.21	.2195	.46	.4771	.71	.7265	.96	.9635
.22	.2299	.47	.4872	.72	.7363	.97	.9727
.23	.2403	.48	.4973	.73	.7460	.98	.9818
.24	.2507	.49	.5075	.74	.7557	.99	.9909
.25	.2611	.50	.5176	.75	.7654	1.00	1.0000

[1] From Rugg's *Statistical Methods Applied to Education.*

INDEX